Ks Docket & issue to the library 15/11/79

VOLUME 86 / NUMBER 1 JANUARY 1979

Psychological Bulletin

(Continued on inside back cover)

ISHED BIMONTHLY BY THE AMERICAN PSYCHOLOGICAL ASSOCIATION

The *Psychological Bulletin* publishes evaluative reviews and interpretations of substantive and methodological issues in the psychological research literature. The Journal reports original research only when it illustrates some methodological problem or issue. Discussions of methodological issues should be aimed at the solution of some particular research problem on psychology, but should be of sufficient breadth to interest a wide readership among psychologists; articles of a more specialized nature can be directed to the various statistical, psychometric, and methodological journals. The *Bulletin* does not publish original theoretical articles; these should be submitted to the *Psychological Review.*

Abstracts: All articles must be preceded by an abstract of 100–175 words. Detailed instructions for preparation of abstracts appear in the *Publication Manual of the American Psychological Association* (2nd ed.), or they may be obtained from the Editor or from APA Central Office.

Blind review: Because reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the author's identity or affiliation should be on separate pages. Every effort should be made to see that the manuscript itself contains no clues to the author's identity.

Manuscripts: Submit manuscripts in triplicate to the Editor, R. J. Herrnstein, *Psychological Bulletin,* Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, according to instructions provided below.

Instructions to Authors: Authors should follow the directions given in the *Publication Manual of the American Psychological Association* (2nd ed.). Instructions on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors are requested to refer to the "Guidelines for Nonsexist Language in APA Journals" (Publication Manual Change Sheet 2, *American Psychologist,* June 1977, pp. 487–494) before submitting manuscripts to this journal. All manuscripts should be submitted in duplicate and both copies should be clear, readable, and on paper of good quality. Dittoed copies are not acceptable and will not be considered. Authors are cautioned to carefully check the typing of the final copy and to retain a copy of the manuscript to guard against loss in the mail.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1979: $40 domestic, $42 foreign, $7 single issue. APA member rate: $15. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Dept., APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in January) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D.C., and at additional mailing offices.

Vol. 86, No. 1 January 1979

Psychological Bulletin

Criminal Behavior of Discharged Mental Patients: A Critical Appraisal of the Research

Judith Godwin Rabkin
Epidemiology of Mental Disorders Research Unit
Psychiatric Institute, New York, New York
New York State Office of Mental Health

Continuing debate on the dangerousness of mental patients between mental health representatives and community members justifies a critical appraisal of the available evidence. Included for review are epidemiological, prospective studies of arrests and convictions among discharged mental patients in comparison to arrests and convictions of the general public and their change over time. Patients with arrest records prior to hospitalization were found to have arrest rates after discharge that far exceeded those of the general public or of other patients. As the number of patients with prior records has increased over time, postdischarge rates for patients considered as a single group have increased accordingly, although patients without prior records continue to have postdischarge arrest rates equal to or lower than those of the general public.

For at least 25 years, investigators of public attitudes toward mental illness have reported that most people dislike, distrust, and fear the mentally ill. Although it has become less socially acceptable to acknowledge such attitudes today, they continue to prevail (Rabkin, 1974, 1976). When these attitude studies were first conducted during the late 1940s and early 1950s, interest was motivated by the desire of mental health professionals to encourage troubled people to seek psychiatric help earlier in the course of their difficulties and to ease the reentry into the community of patients released from distant mental hospitals. More recently, as state and local mental health agencies have accepted the tenets of the community mental health movement and have adopted policies promoting community-based treatment, negative public attitudes about mental illness and mental patients have served as a real and persistent obstacle to the fulfillment of these goals. Communities resist the placement in hotels and boarding houses of older chronic patients discharged after years of hospitalization, often on the grounds that such ex-patients are unsupervised, uncontrolled, and unsuitable neighbors. At the same time, local treatment facilities for both chronic and younger acute patients are resisted for fear that additional mentally ill people will be attracted to the neighborhood, thus compounding the situation. Such resistance has been vocal, effective, and widespread, leading to the passage of municipal ordinances and legal barriers to the establishment of local facilities (Aviram & Segal, 1973; Segal &

Requests for reprints should be sent to Judith Godwin Rabkin, Epidemiology of Mental Disorders Research Unit, Psychiatric Institute, 722 West 168th Street, New York, New York 10032.

Avirām, 1976). It has been estimated that "for every community program that is established and continues to operate, another has been prohibited or closed because of community opposition" (Piasecki, Note 1). Clearly, community opposition continues to be a major problem for those charged with the task of opening new community facilities to treat the mentally ill and retarded.

Investigators have defined two major sources of apprehensiveness that seem largely to account for the pervasiveness and tenacity of public concerns: People perceive mental patients as both unpredictable and dangerous. Unpredictable behavior is always unsettling to social groups because order and stability are threatened. When people are identified as unpredictable, they lose credibility and social standing, and others eventually try to avoid them (Cumming & Cumming, 1965; Nunnally, 1961). However, the most influential factor behind negative attitudes is the perception of the mental patient as dangerous (Cocozza & Steadman, Note 2). This is the cornerstone of public apprehensions and a crucial stumbling block in communications between members of the public and representatives for mental health groups.

It is widely believed by community members that mental patients are likely to display impulsive, violent, assaultive, and otherwise socially disruptive behavior. In public meetings held to consider the applications of new psychiatric facilities in the neighborhood, many community speakers openly express such fears. The vast majority of mental health professionals respond by declaring that such fears are groundless, that mental patients are actually less likely to commit crimes than other people, and that the local opponents are speaking from prejudice rather than fact. Indeed, in such settings, speakers for the mental health establishment have become quite eloquent in condemning the uncharitable and reactionary attitudes of the communities that have expressed resistance to new psychiatric facilities.

Most mental health professionals do believe that mental patients have relatively few encounters with the law and that on the infrequent occasions when they are arrested, it is for the minor offenses such as vagrancy, loitering, or public intoxication. It is furthermore assumed that such charges stem largely from socially inept and unacceptable behavior like urinating on lampposts or wandering aimlessly in the street rather than from purposive criminal acts such as robbery. Psychiatric textbooks, general medical magazines (Farnsworth, 1977), and psychiatric reviews (Gulevich & Bourne, 1970) generally concur in their evaluation of the scientific literature on the criminal behavior of the mentally ill: This literature is regarded as sparse and inconsistent but is overall supportive of the position that patients commit fewer crimes than the general population. This seems to be the prevailing belief not only among mental health professionals but also among most educated, liberal, and thoughtful people.

Debate on these conflicting perceptions of the dangerousness of mental patients persists between mental health representatives and members of neighborhoods slated for the opening of psychiatric facilities. Neither side has resorted to examination of the available evidence, although much heat and controversy have been generated. Such a review seems appropriate and timely and is the object of the present undertaking.

The questions to be considered concern the prevalence of arrests among former hospitalized patients and former outpatients treated for psychiatric disorders in comparison to arrest rates of the general public. The separate literature on the prevalence of mental illness among criminals is not included, nor is a review of efforts to predict dangerousness among mental patients, criminal defendants, or the criminally insane. The studies included for review are epidemiological, prospective studies, and each is addressed to one or more of the following questions:

1. Do discharged mental patients currently have higher arrest rates than members of the general population? (a) Have these rates changed over time? (b) What factors have contributed to such changes, if observed?

2. What are the best predictors of postdischarge arrests?

3. What is the association between arrest risk and diagnostic category?

4. Are discharged mental patients more likely to be arrested for certain types of crime?

5. Does hospitalization reduce the probability of recidivism among patients with prior arrest records?

After a brief review of pertinent methodological issues, each study is presented critically in some detail. Studies are classified as *early* or *recent* and are reviewed in chronological order. Because it may be difficult to keep track of the details of several studies simultaneously, two tables have been prepared as an aid. Table 1 outlines the design of each study, and Table 2 presents selected results. Following presentation of the studies, cumulative findings are summarized according to the questions just posed, and the state of the evidence for each is evaluated. Finally, some implications are drawn with respect to the relevance of these findings for efforts to further develop community psychiatry programs.

Methodological Considerations

The question under consideration here is a deceptively simple one: Are formerly hospitalized mental patients engaged in criminal activity with lesser or greater frequency than other people? The appropriate research design is necessarily a prospective one [1] (sometimes called a cohort design): Two groups of people, one with a history of psychiatric hospitalization and one without but otherwise similar, are followed for equal periods of time to obtain for each group counts of police encounters, arrests, convictions, and incarcerations. Because it has been claimed that mentally disturbed citizens are treated differently by the criminal justice system than are other defendants, a retrospective or case-control design in which the histories of convicted offenders are searched for evidence of mental illness is not as effective, although it is considerably simpler and less costly. The current review therefore focuses on prospective studies.

Since arrests for criminal acts are comparatively rare events and because the expected incidence of arrests was unknown at the start, investigators were obliged to study a large number of people at risk. Among the studies reviewed, samples ranged from 310 to nearly 100,000 patients discharged from the same or similar hospitals within a 1- or 2-year time span. Frequency of arrests within a specified follow-up period ranging from $1\frac{1}{2}$ to $5\frac{1}{2}$ years was computed from police records, and then annual arrest rates were compared with figures compiled at the catchment area, county, state, or national level for the general population of similar age. Comparative rates of arrests for patients and the general population served as the basis for conclusions about the relative incidence of criminal activity among discharged mental patients.

A number of methodological problems become apparent in the course of reviewing studies in this field. Some problems, such as lack of equivalence in the demographic and psychiatric characteristics of patient samples, are more or less unavoidable when one considers together any series of studies conducted by different people in different places at different times. Moreover, such heterogeneity may be regarded as an asset in the sense that replicated findings can be generalized to a broader segment of the population of patients. Other difficulties, such as extrapolating rates based on few cases, are common to any epidemiological study of rare events and can be dealt with by collapsing categories to enlarge the number of cases used to generate rates and by seeking patterns rather than focusing on separate rates by specific category.

Apart from such general considerations, there are a number of problems specific to the field under study that warrant mention.

[1] The term *prospective* is used here in the sense advocated by Lilienfeld (1976) in which two groups, one with and one without a certain characteristic, are followed and rates of a subsequent disorder or event are computed for each. In this case, the two groups are hospitalized mental patients and some subset of the general public, and the events in question are arrest rates within a given follow-up period.

Table 1
Research Design of Studies of Arrest Rates of Discharged Mental Patients: Prevalence Period, Crime Categories, and Sample and Control Group Characteristics

Authors	Years of discharge	Prevalence period: Prior arrests	Prevalence period: Follow-up	Crime categories	*N*	Criteria for inclusion	Control population
Ashley (1922)	1912–1922	—	1 month to 1 year	All arrests	1,000	Patients on parole status from one state hospital	None
Pollock (1938)	1937	—	1 year	All arrests	5,833	All patients on parole status in 1937 from state hospitals	1937 arrests in New York State, over age 14
Cohen & Freeman (1945)	1940–1944	—	2 years	All arrests	1,676	Patients paroled or discharged between 1940–44 from state hospital	1942–1943 arrests for Connecticut residents
Brill & Malzberg (1962)	Fiscal 1947	Lifetime	5 years	All arrests	10,247	Male patients over age 15 discharged from New York State hospitals	1947 arrests for males over age 15 in New York State
Rappeport & Lassen (1965, 1966)	1947 and 1957	5 years	5 years	Five violent crimes only	1,401 in 1947 4,281 in 1957	Discharges from all hospitals (private, public, federal) in Maryland, over age 15	Maryland male and female population over age 15 for five crime categories
Giovannoni & Gurel (1967)	1956	—	4 years	All arrests and all encounters with police	1,142	Male VA hospital patients (95% schizophrenic) under age 60	U.S. residents in cities with over 25,000 people

Table 1 (*continued*)

Authors	Years of discharge	Prevalence period: Prior arrests	Prevalence period: Follow-up	Crime categories	*N*	Criteria for inclusion	Control population
Sosowsky (Note 3)	1966–1970	—	1–5 years	Homicide and assault convictions	99,361 hospital patients; 143,322 outpatients	California patients over age 15 receiving state mental health funds	California nontreated residents over age 15
Sosowsky (Note 3)	1972–1973	6 years	Not given	All arrests, convictions for homicide and assault	301	San Mateo residents over 15 discharged from state hospital (73% schizophrenic)	San Mateo County nonhospitalized residents over age 17
Zitrin, Hardesty, Burdock, & Drossman (1976)	1969–1971	2 years	2 years	All arrests	867	Residence in Bellevue catchment area, about half schizophrenic, aged 10–80 (most 20–50)	Catchment area population and all U.S. cities
Durbin, Pasewark, & Albers (1977)	1969	5 years	5 years	All arrests	461	All state hopsital patients aged 18–64 admitted in 1969 (70% alcoholics)	Wyoming population aged 18–64
Steadman, Melick, & Cocozza (Note 5)	1968 and 1975	Lifetime	19 months	All arrests	1,920 in 1968 1,938 in 1975	Systematic samples of discharged state hospital patients over age 17	New York State population over age 17

Note. VA refers to the Veterans' Administration.

One issue concerns the differential probability that mentally disturbed and other people will be arrested, sent to jail, convicted, and incarcerated. (Studies in this review excluded from consideration those defendants who were hospitalized by court order for psychiatric observation, as well as those classified as criminally insane.) It has been argued that convictions and even arrests are inadequate measures of crime among mental patients because the seriously disturbed defendant and those with histories of hospitalization are rehospitalized instead of arrested. Recent evidence supports this contention. Both Levine (1970) and Lagos, Perlmutter, and Saexinger (1977) found extremely high rates of violent or illegal behavior presented as part of the basis for hospitalization in the admission notes of randomly selected inpatients. Levine reported that 71 of 100 patients had com-

Table 2
Results of Studies of Arrest Rates of Discharged Mental Patients: Samples, Comparative Rates, Overrepresented Diagnostic Groups, and Most Common Offenses

Authors	Diagnostic and sex distribution of sample	Total arrests per 1,000: Patients	Total arrests per 1,000: General public	Arrests per 1,000 for assaults and homicides only: Patients	Arrests per 1,000 for assaults and homicides only: General public	Overrepresented diagnostic groups	Most common offenses
Ashley (1922)		37					
Pollock (1938)		6.9	98.5				
Cohen & Freeman (1945)		26					
Brill & Malzberg (1962)	Males only	12	49	1.3	.8	Alcoholics, drug abusers, psychopathic personalities	All felonies; assaults and homicides
Rappeport & Lassen (1965)	Males only						Robbery
Rappeport & Lassen (1966)	Females only						Aggravated assault
Giovannoni & Gurel (1967)	Males only; 95% chronic schizophrenics			3.27	.94		Violent crimes (27% of all arrests); drunkenness (23% of all arrests)
Sosowsky (Note 3)	Statewide sample			2.96[a]	.43[a]		
Sosowsky (Note 3)	San Mateo County sample; 73% schizophrenics			27.8 11.7[a]	1.8 .41[a]		Violent crimes (17% of all arrests)
Zitrin, Hardesty, Burdock, & Drossman (1976)	47% schizophrenics			7.3	1.59 (U.S.) 6.26 (local)	Alcoholics and addicts (schizophrenics had excess rates for violent crime with bodily harm but not for total arrests)	
Durbin, Pasewark, & Albers (1977)	Male rates only; 69% alcoholics			3.0[b]	1.6	Addicts and personality disorders	Drug offenses
Steadman, Melick, & Cocozza (Note 5)	1968 discharge sample; 27% schizophrenics	73.5	27.5	5.6	2.3	Personality disorders and substance abusers	Drug, property, and violent crimes
Steadman, Melick, & Cocozza (Note 5)	1975 discharge sample; 32% schizophrenics	98.5	32.5	12.0	3.6	Paranoid states and personality disorders	Property, sex, and violent crimes

[a] Conviction rate.
[b] By arrests, not persons.

mitted illegal acts, including 23 acts judged to be felonies, in the course of episodes leading to their hospitalization, none of which were prosecuted. Similarly, Lagos and his colleagues found that 36% of 321 patients manifested some form of violent behavior in the episode leading to hospitalization, of whom only 2.6% were prosecuted. Despite validity problems in these data sources, including the possibility that admitting personnel unduly emphasize violence to justify the decision to admit, it does seem that many dangerous, violent, and illegal acts committed by distressed and distraught people lead to hospitalization rather than arrest. As a result, arrest rates may underestimate frequency of violence among the mentally ill.

Another source of error in use of arrest and conviction rates to represent frequency and type of illegal behavior of the mentally ill concerns the response of the criminal justice system. If indeed charged instead of hospitalized, the disturbed defendant may be acquitted on grounds of insanity. Further, the charge may be reduced. In some states, defendants accused of serious crimes (felonies) cannot be committed to a psychiatric hospital except for brief observation if an insanity plea is under consideration. A common solution is the reduction of a charge to a lesser offense (a misdemeanor) to enable civil commitment of a disturbed defendant (Paull & Malek, 1974). This procedure applies only to psychotic individuals; sociopathy, alcoholism, and drug abuse are not legally acceptable grounds for either acquittal or reduction of charges for commitment purposes. One result is the lowering of overall arrest rates for the mentally ill. Another is the reduction of the severity of charges pressed against schizophrenics, who then appear less prone to criminal acts than nonpsychotic diagnostic groups.

In an effort to obtain less biased measures of criminal activity of mental patients, one may seek to enumerate all police contacts rather than just those culminating in arrests. Apart from the practical difficulties of obtaining such information, however, one cannot equate police contact with criminal activity. First, most police contacts do not lead to arrests. The Federal Bureau of Investigation (1976) indicated that only 21% of index crimes[2] were cleared during 1976. A crime is *cleared* when law enforcement agencies have identified the offender, have sufficient evidence to charge him or her, and have actually taken him or her into custody. Clearance rates vary widely by type of offense, ranging from 79% of murder offenses to 14% of motor vehicle thefts. Police clear a high percentage of crimes against the person, both because of the more intense investigative effort made and because of the greater availability of witnesses to identify the perpetrator and to testify against him or her. In their review of dispositions of encounters between discharged mental patients and the police, Giovannoni and Gurel (1967) also found that only a minority of police encounters culminated in arrests. Although such low clearance rates reflect to some extent limitations of law enforcement resources and procedural constraints, they are necessarily less than 100% because several people may be suspected in the same case and only one is ultimately charged.

Another objection to the use of police contacts as an index of criminal activity is the possibility of differential "whistle blowing" behavior for ex-mental patients and other citizens. Former patients may be more likely to be under police surveillance because of their known status; their social ineptnesss may heighten their visibility, or unfriendly neighbors may call the police with complaints for purposes of harassment. A more basic difficulty is the definition of criminal activity, which is necessarily determined only by a judge in court, so that police contacts may be interesting to study, but are insufficient evidence in themselves.

A similar argument may be made with respect to arrest records. Here also, the de-

---

[2] Index crimes include seven offenses selected because of their seriousness, frequency of occurrence, and likelihood of being reported to the police. They are murder, forcible rape, robbery, aggravated assault, burglary, larceny–theft, and motor vehicle theft. The Federal Bureau of Investigation's annual national reports refer to these seven crimes only.

fendant may be innocent. However, it seems likely that a sufficient number of mental patients are taken out of court channels before completion of a trial and passage of a verdict so that use of conviction rates alone gives a misleadingly low estimate of the criminal activity of the mentally ill (Zitrin, Hardesty, Burdock, & Drossman, 1976). In all of the studies reviewed here, with one exception, measures of criminal behavior were based on arrest rates. The exception, a pair of studies conducted by Sosowsky (Note 3), found conviction and incarceration rates highly correlated with arrest rates.

Another problem concerns computation of crime rates by specific category. There is no national criminal code, and definitions and classifications of offenses vary between jurisdictions and over time within a given jurisdiction. A particular offense may be a felony in one state and a misdemeanor in another, or it may be reclassified as one or the other 5 or 10 years later. It is therefore difficult to compare arrest rates, for example, for concealed weapons between Wyoming and New York. One solution is to consider broader categories of crime rather than specific offenses, although here the problem becomes what system of classification to use in developing such categories. It may be to circumvent such difficulties that the majority of recent investigators chose to emphasize the relative occurrence of violent crimes, defined as homicides plus assaults, in which the charges are less ambiguous, the proportion of arrested perpetrators is comparatively high, and the task of classification is less difficult.

The foregoing issues are related to the task of defining and counting cases, the results of which constitute the numerator of a rate. There are also problems in selecting an appropriate denominator, which is defined as all cases at risk for the event in the numerator. Almost all investigators have computed the denominators of their patient rates as the number of discharged patients in their sample. Because of the large sample sizes used in most studies, the size of error in using this denominator is probably minor, especially if the sample is not broken down into subgroups. However, some patients die, move to other states, or spend time away from the community in jails, mental or medical hospitals, nursing homes, or elsewhere. In each instance, they are no longer "at risk" for arrest. In only one study (by Giovannoni & Gurel, 1967) was an effort made to determine the actual population at risk; others overestimated the denominator size and thus obtained lower arrest rates. Another prevalent error is the inclusion of patient arrests in general population figures derived from Federal Bureau of Investigation (FBI) records. Only Sosowsky (Note 3) excluded treatment patients from his population arrest rates.

An additional source of error is lack of genuine equivalence between patient and general population groups. No one has ever claimed that state hospital patients (who constitute the vast majority of patients studied in this field) are representative of the general population. At best, they represent the less fortunate members of society who may be collectively described as lacking social status, financial resources, occupational skills, and often family ties. If mental patients in public facilities were compared to their peers in these terms, each of which is associated with the distribution of both arrests and mental illness in the general population, fewer differences might be found.

Although it is true that studies in this research area are individually and collectively incomplete and in many respects insufficient in their designs and analyses, their review appears warranted on several grounds. First, they represent the only available empirical evidence in an area of tremendous public concern. At the present time intense policy debates are being conducted regarding the relative rights of patients and of communities into which they are discharged. The cumulative evidence that is derived from these studies may help to clarify major elements in these debates. In addition, summarization and criticism of this literature should serve as a guide to future research by indicating important, unresolved questions and suggesting more fruitful methods for their investigation.

In my opinion, the methodological issues raised here are both relevant and significant,

but are probably not major sources of weakness in the group of studies to be reviewed. The direction of bias is not consistent. Although no study is flawless, no problem is universal, so that agreement of results across studies lends robustness to their conclusions, the foregoing difficulties notwithstanding.

Literature Review

Early Studies [3]

Between 1922 and 1955, four studies were conducted regarding the subsequent arrest rates of paroled or discharged mental patients. These studies have been accepted uncritically and quoted extensively during the past 50 years to substantiate the prevalent belief that mental patients are less dangerous and less inclined to criminal activity than other people.

In 1922 the superintendent of the Middletown State Homeopathic Hospital in New York State published a brief report describing the subsequent careers of 1,000 patients paroled from his hospital during the previous 10 years (Ashley, 1922). Ashley noted that "as the parole period up to three years ago covered only a period of from one to six months the data are not as complete as might be desired" (p. 64). This passage is quoted to show the basis for computation of period prevalence rates, which is evidently not entirely clear. The length of parole during the last 3 years of the study was not reported, nor was the distribution of paroles over time. Assuming an equal distribution and using the mean parole duration, one is led to conclude that the report was based on a 3-month follow-up for 700 of the 1,000 cases. Later publications referring to a 1-year parole period serve as the basis for the conjecture that the parole period during the last 3 years of Ashley's study did not exceed 1 year.

Ashley carefully tallied recovery rates, discharge status, economic status, social adjustment, and readmissions of his 1,000 parolees, of whom two thirds were female. Sixty-four percent were either partly or fully restored to their former social roles, including the 46% that were recorded as cured. Although nearly one quarter of the parolees were involved in social conflicts serious enough to be brought to the attention of the hospital authorities, only 12 were arrested. The source of arrest figures is not given, but it seems likely that they were obtained from reports of relatives or neighbors to parole officers. The 12 arrests were for offenses including "vagrancy, assault and battery, forgery, swindlery or profiteering" (p. 65).

Twelve arrests for 1,000 patients during a mean time period of 3 months after hospital release would be equivalent to an annual rate of 37 per 1,000.[4] Another 223 parolees were reported in Ashley's words to have engaged in "antisocial acts either resulting from or conducive to further mental trouble" (p. 65), but they were not arrested. One may conjecture that only 12 of 235 known antisocial acts resulted in arrest because of the accessibility of parole officers as negotiators or problem solvers or because of the reluctance of others to press charges against parolees. Arrest rates for the general population of that time were not reported, so one cannot determine the extent to which the observed rates depart from expected rates.

Two other points regarding Ashley's arrest figures warrant consideration. First, two thirds of his sample were female, and women's arrest rates historically have been much lower than men's. In addition, parole status is not the same as the unconditional discharge of

---

[3] For the reader's convenience and to facilitate comparisons, all rates cited here are expressed as per 1,000 even if the authors used a different basis.

[4] The rate of 37 per 1,000 is derived by estimating that 70% of the sample had a mean follow-up of 3 months and the rest a mean follow-up of 1 year. Thus, 70% of 12 was multipled by 4 to give an annual basis and then added to the remaining 30% ($70\% \times 12 = 8.4 \times 4 = 33.6 + 3.6 = 37.2$). This estimate is vastly greater than the rate computed and reported by Pollock (1938) in his article that summarizes Ashley's (1922) findings and that has since been widely quoted. Pollock interpreted the passage cited above to mean that Ashley's cases "were under observation on the average nearly 5 years," yielding an "average annual rate of arrests . . . of 2.4 per 1,000" (p. 239). Unless Pollock had access to additional information from Ashley other than the 1922 article, which Pollock does not mention, it would seem that his estimate is based on a misreading of Ashley's article.

contemporary patients. Parole could be revoked by the unilateral decision of hospital representatives, so patients may have been more careful about their behavior and more closely supervised by family members than patients described in recent studies.

In 1938, Pollock published a study of the legal offenses committed by patients paroled during fiscal 1937 from all New York State hospitals. At that time, patients considered eligible for release were paroled for 1 year before discharge. "To be deemed eligible for parole, the patient must be harmless, tractable, and able to mingle with others without causing trouble. . . . No patient is placed on parole unless there is a safe and suitable home to receive him and a relative or friend to welcome and care for him" (p. 241). Evidently patients were rigorously screened before their parole in terms of social as well as psychiatric criteria.

Pollock wrote his article to counteract the "frequently made charge that the paroling of mental patients by State hospitals is a dangerous procedure" (p. 236), and his data certainly look convincing although their sources are ambiguous. In 1937, nearly 65,000 patients were distributed among the 20 state facilities, and on an average day there were 5,833 patients on parole, including a slight excess of males. Pollock reported that paroled patients were arrested for 40 offenses, including 29 misdemeanors and 11 felonies; all but 3 of the misdemeanors were committed by men. Dividing 40 by 5,833, he obtained an arrest rate of 6.9 per 1,000. Unfortunately, Pollock did not indicate the source of his arrest data or the time period covered. Since this rate is compared to that of the state's general population in 1937, the implication is that the 40 arrests were all those committed by all patients paroled during fiscal 1937. Accordingly, this rate is substantially lower than that of the general public over 14 years of age in New York State, which was 98.5 per 1,000 in 1937[5] or 14 times higher than the arrest rate of paroled patients. As Pollock put it, whereas 40 arrests were made among parolees, 576 arrests were made among an average normal group of like size in the general population.

Pollock cited, as further evidence of the arrest rate of parolees, a 10-year follow-up study of 741 patients paroled from two New York State hospitals conducted under the direction of R. Fuller and apparently never published. In his one-paragraph summary of the findings, Pollock reported that "among the 289 male patients there were 19 arrests and among the 452 female patients five arrests. The annual rate of arrests among males was 0.7% and among females 0.1%" (p. 240). No further information was provided.

Pollock attributed these low arrest rates to two factors: the extreme care exercised in screening patients for parole and their constant surveillance while on parole. Not only did social workers regularly visit the patients to interview them as well as their families but patients were required to attend parole clinics at specified intervals. When the parole officer thought it necessary, the patient was fetched by the hospital car and promptly readmitted. Pollock concluded that "the hospital methods are well planned and are producing beneficial results to the patients and communities served" (p. 242). The only reason to entertain some doubt about his figures, which as indicated are undocumented, is his treatment of Ashley's annual arrest rates (discussed earlier).

Like Ashley and Pollock before them, Cohen and Freeman (1945) were associated with a state hospital and were distressed by public resistance to the presence of discharged mental patients in the community. They were eager to promote patient acceptance by showing that they were not dangerous and toward this end conducted a follow-up study of 1,676 patients who were paroled and discharged from one of three state hospitals in Connecticut between 1940 and 1944. (This is the only early study not based on New York State patients.) Patients were followed for an average period of 2 years during which 87 or 5.2% were arrested, yielding an annual arrest rate of 26

[5] This is a peculiarly high figure compared to general population arrest rates subsequently reported (1947 through 1975), which are never more than half as high.

per 1,000. It occurred to Cohen and Freeman that patients who were in trouble with the law after discharge might also have police records before their hospitalization. They found that 314 or 18.4% of their patients had such prior records, and all but 6 of the patients arrested after discharge came from this group. In other words, patients without prior arrest records were almost never in legal difficulty after hospital discharge, whereas 93% of those subsequently arrested had records. Both before and after hospitalization, more than two thirds of arrests were for "drunkenness and breach of peace." Excluding sex offenses for which no population rates were available, the total annual felony rate per 1,000 for discharged patients was .9, compared to a rate of 13.7 per 1,000 for the general population. From these figures Cohen and Freeman concluded that "these patients who have left mental hospitals are not as dangerous to the community as those who have never been judged mentally ill" (p. 699).

Cohen and Freeman considered the etiological significance of mental status in the production of criminal behavior but reached no firm conclusions. They suggested that some arrests seem related to mental illness and that "once the patient is hospitalized and released, the arrest-precipitating behavior is either no longer present or is under some social control," (p. 699) referring in the latter point to control by other people. On the other hand, they felt that many of their recidivists were "inadequately adjusted for reasons which are not directly related to their mental illness at all. . . . [They] seem merely to have been following old anti-social patterns" (p. 700). In subsequent literature reviews such as that by Rappeport, Lassen, and Hay (1967), their idea that hospitalization produces a reduction in criminality is emphasized, but the notion of coexisting patterns of mentally ill and antisocial behavior is overlooked.

The next, and perhaps most influential, study in this series of early investigations was conducted for the New York State Department of Mental Hygiene by Brill and Malzberg (1962). Their report, based on state hospital discharges during fiscal 1947, was circulated within the department in 1954 but was not otherwise disseminated until 1962. Even then, it was distributed in the form of a supplementary mailing by the American Psychiatric Association rather than published in a standard journal. Nonetheless, it became widely known and often cited, both because of its large scope and because of its investigation of historical variables associated with arrest rates as well as of the rates themselves.

In their work, Brill and Malzberg studied the arrest rates of 10,247 male patients over age 15 released from New York State mental hospitals during fiscal 1947. Arrests at any time before and for 5½ years after hospital release were traced by means of a fingerprint registry in Albany for approximately half the sample.[6] Arrest rates for patients were compared with those of all males over age 15 in New York State for 1947.

Patients were divided into two categories: those with a prior history of arrests and those without. Arrest rates after hospital release were computed for both groups separately and together. The patients with prior arrest records (15%) turned out to have dramatically higher subsequent arrest rates than the other patients as well as than the general population, whereas the patients without arrest histories (85%) had considerably lower rates than the general population for misdemeanors, all felonies, crimes of violence, and all crime categories combined. When the patients were considered together as a single group, their felony arrest rates were higher and their total arrest rates were lower than those of the general population.

As seen in Table 3, the patients fall into two very different groups depending on arrest history. As a group, patients with prior arrests had subsequent arrest rates ranging from 6 to 8 times higher than those of the general population and 9 to 16 times higher than those of the other patients. Although

[6] These patients ($N = 5{,}354$) were admitted after fingerprinting became a routine part of the admissions procedure. Arrest rates for the remaining 4,893 were calculated based on demographic and historical characteristics.

Table 3
Annual Arrest Rates per 1,000 for Male Patients Released From New York State Hospitals and for the General State Population in 1947

Offense	Patients with arrest records	Patients without arrest records	Total patients	General population
Misdemeanors	32.8	2.0	6.7	45.8
Felonies (all)	26.3	1.6	5.5	3.3
Homicides and assaults only	5.3	.6	1.3	.8
All categories	60.0	3.6	12.2	49.1

Note. These data are from Brill and Malzberg (1962).

constituting only 15% of the patient group, their excess of arrests in the felony category, especially for violent crimes (homicide and assault), was sufficient to influence the arrest rate of all patients combined compared to that of the general population. This excess was due to only one third of the patients with prior arrests; 66% of these patients and 98% of the other patients were not arrested after hospital release.

A critical issue that Brill and Malzberg made some effort to address, which has not been dealt with by other investigators either before or since, is whether the recidivism rate among ex-convicts who have been in psychiatric hospitals is the same, lower, or higher than that of people with arrest histories who have not been hospitalized. Based on information from unnamed "expert criminologists," Brill and Malzberg reported an overall annual rearrest rate of 5% among paroled ex-convicts, which is similar to the 6% rate among released mental patients with arrest records. The unnamed experts also reported that only 5% of the general population in 1947 had arrest records, in contrast to 15% of the patient population.

Brill and Malzberg also looked into the demographic and psychiatric characteristics of patients in relation to arrest patterns. Overall, the same factors associated with recidivism in the general population were found to characterize rearrested patients, including unmarried status, youth, alcoholism, drug addiction, and residence in delinquency areas. Rearrested patients were overrepresented in the diagnostic categories of alcoholism, drug abuse, and psychopathic personality and significantly underrepresented among the major functional psychiatric disorders. Symptom severity and length of hospital stay were negatively correlated with recidivism. As one might expect, lower arrest rates prevailed among voluntary admissions. Brill and Malzberg concluded that "statistically the group of previously arrested patients behaves more like a segment of the correctional population than a primarily psychiatric one" (p. 7).

In summary, their data show that mental illness and psychiatric hospitalization do not raise the probability of subsequent arrest above that existing before hospitalization and do not create such a tendency if it did not previously exist. This is an observation of major and enduring significance, and it is perhaps ultimately all that must be said on the subject of crime and mental illness. Their data demonstrate that crime rates for ex-patients without arrest records are lower than those for the general population and are inflated by the presence of patients with police records. In 1947, and as one sees even more so today, men with arrest records constitute a larger proportion of the patient population than of the general population, and it is this that contributes to the higher arrest rates of mental patients when they are considered as one group.

The preceding four studies (sometimes counted as five when the paragraph about Fuller's findings, cited in Pollock, 1938, is considered separately) together constitute the foundation for psychiatric reassurances that former mental patients are no threat or danger to their neighbors. In fact, the first two studies concur in their results and interpretations, whereas the third and fourth identify the major factor contributing to higher arrests

among discharged patients, providing a framework within which to integrate the apparently contradictory findings of studies published before and after 1965.

The first three studies are difficult to evaluate because so little information is provided with respect to record sources, case representativeness and selection, and computation of rates. Nevertheless, even if the authors minimized patient arrests and used high estimates of population arrest rates to yield an exaggerated contrast, it is reasonably certain that mental patients discharged before World War II were less often arrested than were members of the general public.

Recent Studies

Since 1965, eight American studies including nine samples were designed to contribute further empirical evidence to the question of the dangerousness of discharged mental patients. Each study found that arrest or conviction rates of former mental patients equaled or exceeded those of the general population in at least some crime categories when patients were considered as a homogeneous group. Although each study has some limitations and all do not provide equivalent data, they are cumulatively persuasive in their evidence that patterns of arrests among former mental patients are very different from and far higher than those reported earlier, both in absolute terms and in comparison to the general public.

In terms of temporal sequence, the studies of Rappeport and Lassen (1965, 1966) serve as a bridge between the earlier studies and contemporary ones, since they investigated arrest rates of patients discharged during the fiscal years of 1947 and 1957 from all Maryland psychiatric hospitals. Findings for men and women were published separately (in 1965 and 1966, respectively). Rappeport and Lassen used a data set superior to data sets of other studies in that it included discharges from federal and private as well as state facilities. The total sample consisted of 708 men and 693 women in 1947 and 2,152 men and 2,129 women in 1957. Arrest records for 5 years preceding and 5 years following hospital discharge in 1947 and 1957 were collected from each police jurisdiction in the state. Comparative arrest rates for the state population for equivalent time periods were obtained from the FBI for males and females over age 15 separately.

Arrests were recorded for only five offenses, all of which were violent crimes against persons: murder, negligent manslaughter, rape, robbery, and aggravated assault. The obtained arrest rates are thus not comparable to *total* arrest rates reported in other studies. Further, they counted arrests only in the 5 years preceding this hospitalization rather than during the lifetime of the patient, as in the Brill and Malzberg (1962) study.

For each of the five crime categories, patient rates were equivalent to or higher than those of all Maryland residents. Male rates for robbery and female rates for aggravated assault were particularly high. Arrest patterns were not notably different between 1947 and 1957 for either males or females, despite expectations that newer treatment techniques would have reduced posthospitalization arrests. The authors drew two major conclusions. First, they found no support for the contention that mental patients "are to any great extent less involved in criminal behavior than those in the general community" (Rappeport & Lassen, 1965, p. 779). Second, they concluded, as did Brill and Malzberg, that recidivism among patients with histories of prior arrests is "not unlike that seen in the general community" (p. 779). In summary, they observed that "we as psychiatrists may be biased when we malign others for suggesting that some of our patients represent a threat to the community" (p. 779).

Rappeport and Lassen attempted to analyze arrests by diagnostic category; thus for each of the four time periods studied (before and after 1947 discharges and before and after 1957 discharges) they tallied arrests by diagnostic category. The two categories that accounted for the largest proportion of arrests in all four time periods were those of alcohol intoxication and schizophrenia. Since Rappeport and Lassen did not say how diagnoses were distributed in the entire sample, one cannot determine whether these diagnoses are underrepresented or overrepresented in the

subset of patients who were arrested. Further, only percentages were reported, which can be grossly misleading, as in the breakdown for females arrested after their 1947 release. The authors showed in tabular form that 33% of arrested females were diagnosed as manic-depressive, 33% as neurotic, and 33% as mental defective. In the footnote of another table they stated that only three arrests were made of female patients during this time period, so that the diagnostic categories each contain a single case.

These lacunae are partly offset by the presentation of arrest rates per 100 within diagnostic category for each of the four time periods. In these tables and in the text the misleading effects of the previous presentation of percent arrests by diagnosis are diminished, as one can see that the arrest rates per 100 for both alcohol intoxication and schizophrenia no longer appear outstanding, and that in fact no consistent association between arrest rate and diagnosis emerges across the four time periods. Some effort at statistical analysis would have been helpful in this context. As presented, this material may have some utility in defining patterns within the data set, but it is not otherwise informative.

Analyses by type of crime are also of doubtful value in Rappeport and Lassen's studies (1965, 1966), since in several categories the number of events was so few that the reliability of computed rates must seriously be questioned. For example, no more than two arrests for murder by male patients were recorded in any single time period; consequently, estimation of rates per 100,000 based on such rare occurrences is misleading because of the high probability of error. However, the cumulative rates for all five crime categories combined are based on sufficient numbers to be relatively stable, and the authors' conclusions seem reasonable and justifiable.

The next study, by Giovannoni and Gurel (1967), is distinguished by the national distribution of the patients studied, its exceptionally comprehensive pursuit of patients' status at follow-up, its restriction to the single diagnostic category of chronic schizophrenia among males, and its analysis of all police contacts and their dispositions, not only those that ended in arrests. Unlike any other researchers in this field, Giovannoni and Gurel computed arrest rate denominators by meticulously measuring "in-community days" and thus precisely determining the number of patients at risk for arrest in a given period of time. Unfortunately, criminal records preceding hospitalization were not reviewed, so that recidivism, which others have regarded as a major component of postdischarge arrest rates, could not be measured.

Giovannoni and Gurel studied the subsequent arrest records of 1,142 males under 60 years of age, 95% of whom were chronic schizophrenics who were released from Veterans' Administration hospitals in 12 states during 1956 and who remained out of the hospital for at least 30 consecutive nights. For each patient, days spent in the community were computed by subtracting from the total those days spent in penal or psychiatric facilities. Follow-up information was gathered from periodic interviews as well as from institutional records. Outcome data consisted of all "socially hazardous" incidents in which patients were involved with police and not only those leading to arrest, since the authors were interested in seeing how often ex-patients who violated the law were rehospitalized rather than arrested.

During the 4-year follow-up, 156 of the 1,142 patients were involved in 192 offenses. In 48% of the incidents, the patient was sent to jail; of these patients, 40% were rehospitalized either directly or after a brief stay in jail. In 12% of the incidents other actions were taken including fines, probation, and dismissal. In the absence of a similar analysis of incident disposition of nonpatient police encounters, one cannot be sure that use of arrest rates alone underestimates the number of incidents in which patients were involved, but these figures suggest such a likelihood.

Offenses were analyzed by type, and rates for each type were compared to an average of national rates from 1957 to 1960 for cities with over 25,000 people, obtained from FBI records. It was necessary to use national figures despite their known limitations because the patients had distributed themselves

Table 4
Conviction Rates per 1,000 for Violent Crimes in California in 1971

Offense	Former state hospital patients	Former outpatients in county program	Nontreated state population over age 15
Assault	1.96	2.57	.33
Homicide	.98	.20	.10
Total no.	99,361	143,322	13,982,155

Note. These data are from Sosowsky (Note 3).

throughout the country. When offenses were grouped into crimes against persons, property, and morals, crimes against persons accounted for twice as many incidents (27%) as did crimes against property (12%). Another common offense was drunkenness. This is not surprising because two thirds of those arrested were rated as problem drinkers by interviewers. Since one is not told how many of the patients who were not arrested were also rated as problem drinkers, it is not clear whether problem drinking is a useful predictor of arrests in this chronically ill population; but in Giovannoni and Gurel's (1967) study, problem drinking seems to be a common characteristic of patients who were arrested.

Estimated annual rates of offense were computed by crime category for patients and compared to general population rates. Despite the low average annual number in each category (ranging from zero to two), which raises serious questions about the reliability of each of the rates, an overall pattern of differences between patient and general population rates emerged that seems more substantial and suggests true underlying differences. Patient rates were higher (compared to the general population) for several violent crimes against persons (homicide, aggravated assault, and robbery but not forcible rape) and much lower for crimes against property (petit and grand larceny, burglary, and auto theft). For example, the estimated annual rates per 1,000 for homicide were .99 for patients and .05 for the general population, whereas burglary rates for patients were .65, compared to 5.1 for the general population.

These findings are of interest first because the patients were chronic schizophrenics, a category infrequently associated with arrests and convictions in other studies. Second, the patients were discharged from 12 different hospitals that varied in the rigor of their discharge criteria among other things, and they were arrested under the laws of many states, ruling out these two factors as possible explanations for the obtained results. These findings are not congruent with those of Rappeport and Lassen (1965, 1966) concerning arrest rates by specific charge within the category of violent crime, but both studies show cumulatively high arrest rates for crimes against persons committed by discharged mental patients compared to the general population.

The California State Department of Health sponsored two studies of crime and mental illness that were conducted by Sosowsky (Note 3). He first undertook a statewide study of arrest, conviction, and incarceration rates for violent crimes (homicide and assault) of former state hospital patients and outpatients at state-supported facilities compared to the nontreated population of California. In the second study he reviewed the criminal records of a cohort of state hospital patients in San Mateo County (California) and compared their arrest and conviction rates with those of the nonhospitalized county population in 1974.

In the first study, Sosowsky listed all patients over age 15 who received California mental health funds (inpatient and outpatient) between 1966 and 1970 and matched them against the names of those convicted in the state for homicide and assault in 1971. Conviction rates were computed as shown in Table 4.

As the author noted, these rates did not take into account ex-patients arrested and ex-offenders treated in other states or those patients not at risk for becoming offenders because of death or movement out of state, but in view of his sample sizes these factors are probably insignificant. His results show that former patients are convicted at a significantly higher rate than are the nontreated population. Combining the two categories of offense, ex-hospitalized patients' rates were seven times and outpatient rates six times greater than those of nontreated California residents. Analysis by age and sex showed this excess to be true for all age groups and both sexes, especially for those under 20 and over 50. Female excess was even greater than that of males compared to the nontreated general population, although male rates for both patients and nonpatients were higher than equivalent female rates in absolute terms.

In his second study, Sosowsky analyzed arrests and convictions of 301 San Mateo County residents admitted to state hospitals between June 1972 and December 1973. The sample included over 95% of all patients that lived in San Mateo County who were admitted to state psychiatric facilities during this period. State criminal justice records were searched for all arrests and convictions of these patients from 1966 through 1973. It seems likely that most arrests preceded hospitalization, since the admission period fell at the end of the 8-year criminal record survey. In fact, 28 patients were still hospitalized when data analysis was begun in March 1974.

Patient arrest and conviction rates were computed for all offenses and for violent and nonviolent offenses separately. Violent offenses included those involving direct bodily harm (murder, assault, and rape) and potential for harm (robbery, burglary, crimes against children, threatening violence, possession of burglar's tools or weapons, kidnapping, arson, and rioting). Nonviolent offenses included all misdemeanors except violations of the motor vehicle code. Patient arrest and conviction rates were compared to those of the nonhospitalized San Mateo County population during the same time period. This is the only study in which the nontreated rather than the total—nontreated plus treated—population was used as a comparison group.

Of the hospitalized patients studied, 67% were male, most were between the ages of 20 and 39, 21% were nonwhite, and 73% were diagnosed as schizophrenic. Compared to the county's general population, males, nonwhites, and those in the 20–40-year-old age range were considerably overrepresented. A total of 47% of the patients were arrested at least once between 1966 and 1974, either before, after, or both before and after hospitalization. On the average, each patient was arrested 3.6 times over the period surveyed. Of the total group, 23.6% were charged with a violent offense and 23.6% were charged with a nonviolent offense (misdemeanor) only.

Since the variables of age, sex, and ethnicity are each independently associated with higher arrest rates in the general population according to FBI findings (e.g., Federal Bureau of Investigation, 1976) and since the patients were disproportionately young, male, and black, the sample was predisposed toward higher arrest rates than those of the general public. However, Sosowsky analyzed arrest rates for violent crimes within these categories, and patients continued to have much higher rates than their age, sex, and ethnic counterparts in the general population. For example, black male patients had an arrest rate of 6%, compared to 2.5% for black males in the county population. Patient excess therefore cannot be accounted for by skewed distributions of demographic variables.

In a subsequent analysis (Sosowsky, 1978) of the second study, Sosowsky raised the issue of the relation between diagnosis and crime category. He presented a table showing that 73% of the sample were schizophrenic, whereas 80% of patients arrested for violent offenses were so diagnosed. In contrast, only 70% of patients arrested for nonviolent offenses were schizophrenic. In the absence of statistical analysis, the very appearance of this table suggests a meaningful relationship between diagnosis and arrest category, but in fact these differences are not statistically significant (using a chi-square, the probability of the first distribution is .16 and that of the second is .45).

After analyzing arrest rates for all offenses combined, Sosowsky (Note 3) compared rates of arrests and convictions for the violent crimes of homicide and assault taken separately, for patients and for the county population. He found an enormous excess among the former. Patient arrests were 15 times higher and conviction rates 29 times higher than equivalent rates for the county population. Computed as rates per 1,000 for arrests and convictions, the arrest rate of patients was 27.80, compared to 1.80 for the county population. The patient conviction rate was 11.70, whereas the county population conviction rate was .41.

Together, Sosowsky's studies (Note 3) offer powerful support for the proposition that both outpatients and hospital patients in state-supported facilities have higher arrest and conviction rates for violent crimes than do nontreated California residents, and presumably, the general public. Unfortunately, these data do not contribute to an understanding of the causal direction between mental illness and crime because Sosowsky did not classify arrests in terms of whether they preceded or followed psychiatric hospitalization. He also did not analyze the recidivism rates of discharged patients with prior arrest histories. Thus, although the magnitudes of the arrest rates he reported for patients are impressive, they are difficult to interpret and they cannot be used to understand the relationships between the criminal justice and mental health systems or between criminal and disordered behavior.

The investigation of patient arrest rates conducted at Bellevue Psychiatric Hospital in New York City by Zitrin et al. (1976) achieved a certain notoriety even before it was published. After its presentation at a psychiatric meeting, a local organization concerned with the rights of mental patients threatened to bring suit against the authors on the grounds that their intent from the start was to discredit patients and to promote their involuntary detention in hospitals.

Zitrin and his colleagues studied 867 male and female patients admitted to this acute-care municipal hospital between July 1969 and July 1971 who lived in the hospital's catchment area (14th to 42nd Streets in Manhattan). Arrest records were located for each patient for the 2 years preceding and following this admission. The patient sample included a slight majority of males and an age range of 10 to 80, although over three quarters were between age 20 and age 50. Multiple hospitalizations were common; the 867 patients had a total of 2,000 admissions to Bellevue during the 2-year study. One patient was admitted 22 times. The authors reported that nearly half of the patients were diagnosed as schizophrenic, 7% as alcoholic, 6% as drug dependent, 8% as neurotic, and 32% as other. However, the diagnostic picture was often blurred, and there were notable overlaps between the categories of schizophrenia and substance abuse. For example, of the 42 patients with a primary diagnosis of schizophrenia who were among the 85 patients arrested for crimes of violence, 20 had a history of drug abuse, 8 had been drug dependent and alcoholic, and 2 others had previously been diagnosed as drug addicts. In other words, three quarters of the patients labeled schizophrenic in this subgroup previously were identified as substance abusers. Similarly, of the 10 alcoholics arrested for violent crimes, 4 had been diagnosed as schizophrenic on at least one previous occasion, and 1 was a drug addict. Evidently, at least among mental patients arrested for violent crimes, the boundaries between these three diagnostic groups (schizophrenia, alcoholism, and drug dependence) are fluid, and computation of comparative incidence rates may be misleading. A similar overlap was identifiable among schizophrenics and substance abusers in this sample who were not subsequently arrested (Zitrin, Note 4).

Overall, 202 or 23% of the patients studied were arrested one or more times during the 4-year period reviewed. Compared to the 45% arrest rate Sosowsky (Note 3) reported for his cohort of state hospital patients, this figure does not seem extreme, except for the fact that Sosowsky recorded arrests during an 8-year period surrounding hospitalization, whereas the Bellevue period of record surveillance was only 4 years.

Patient offenses were classified as violent

Table 5
Age-Adjusted Arrest Rates per 1,000 in 1972 for Violent Crimes

Offense	U.S. cities	Bellevue catchment area	Patient sample
Murder	.15	.45	.30
Aggravated assault	1.44	5.81	7.02
Robbery	1.06	8.81	8.38
Rape	.17	.57	1.15

Note. These data are from Zitrin, Hardesty, Burdock, and Drossman (1976).

crimes involving direct bodily harm (murder, assault, forcible rape, and sexual abuse), violent crimes involving potential for harm (robbery, burglary, weapons possession, possession of burglar's tools, arson, inciting to riot), and nonviolent crimes (presumably all misdemeanors except for simple assault, which is included with other forms of assault as a violent crime). The inclusion of *violent crimes involving potential for harm* in the violent crime category extends its meaning to somewhat less severe offenses than the term as used by others.

Using this classification, 85 or 10% of the 867 patients were arrested for violent crimes and 13.5% for nonviolent crimes. These 85 patients committed a total of 155 crimes of violence, the most common being assault, followed in order of frequency by burglary, weapons possession, and robbery. The age distribution of the violent offenders peaked in the 20–30-year interval, whereas the modal age of nonviolent offenders fell into the 30–40-year interval.

The 85 patients arrested for violent crimes during the 4 years under review were arrested a total of 366 times for various offenses—an average of 4 times each. Twenty-eight of them were arrested only during the 2 years before admission, 21 were arrested only after discharge, and 36 of this subgroup were arrested both before and after this hospital stay. Among the 117 patients arrested for nonviolent offenses, a larger proportion was arrested before the hospitalization than after the hospitalization (Zitrin, Note 4).

Zitrin and his colleagues analyzed crime categories in terms of the four diagnostic groups selected for study. Schizophrenics constituted about half of the sample and accounted for half of all arrests. Differences did emerge in relation to type of crime. Schizophrenics constituted two thirds of patients arrested for violent crimes involving bodily harm (21 out of 30), but were underrepresented among arrests for nonviolent offenses. These results agree with the findings of Giovannoni and Gurel (1967) that schizophrenic patients, when they behave disruptively, are indeed arrested and charged with criminal conduct rather than simply rehospitalized.

Both alcoholics and drug abusers were heavily overrepresented in all crime categories, as Brill and Malzberg (1962) observed earlier. Constituting only 7% and 6% of the patient sample, respectively, they accounted for nearly one third of all arrests. Addicts were more likely to be charged with violent crimes and alcoholics with misdemeanors. Like schizophrenics, patients diagnosed as neurotic were proportionately represented among total arrests and also within crime categories.

The final analysis reported was a comparison of patient arrest rates with those of residents in the same catchment area and with 1972 rates for cities nationally (Federal Bureau of Investigation, 1973), excluding children under age 15 from all groups. As Table 5 shows, arrest rates both for patients and for all residents of the Bellevue catchment area are far higher than those of pooled residents of all American cities. There is comparatively less difference between patients and their catchment area neighbors. The authors did not report the statistical significance of these differences, but it seems from inspection that across the four categories the difference is negligible, although for the offenses of rape and aggravated assault patient rates are higher. Since this catchment area group is most nearly comparable to the patient sample in terms of the social and demographic characteristics of all the studies reviewed, the diminished contrast observed here is noteworthy.

Zitrin et al. (1976) concluded that their data do not support the commonly held belief among mental health professionals that

the mentally ill commit fewer crimes than the general population. They urge provision of more effective aftercare services, together with the introduction of a provisional discharge category in which community stay is contingent on clinic attendance. In agreement with Brill and Malzberg, and Giovannoni and Gurel, they attribute the growing arrest rates among patients to "the increasing diversion of arrested persons from the criminal justice channels to mental hospitals" (p. 147).

A report of arrest rates among Wyoming state hospital patients provides a contrast of setting to the largely urban, industrialized areas previously investigated. Durbin, Pasewark, and Albers (1977) studied the arrest records of all patients aged 18 to 64 (286 men and 175 women) admitted to the only state facility in 1969, excluding patients remanded by criminal order for psychiatric evaluation. This sample was unusual diagnostically in that 63% of the male patients had a primary diagnosis of alcoholism and another 6% had such a secondary diagnosis. Arrest rates for male and female patients separately were derived from police records covering the 5 years before and the 5 years after hospitalization, from 1964 to 1973. Unlike investigators in any other study except that by Giovannoni and Gurel (1967), Durbin used a correction factor in computing arrest rates based on the mean number of days patients were hospitalized during this 10-year period.

Arrest rates of the patient sample were compared with those of the general Wyoming population aged 18 to 64, obtained from state records, for men and women separately. From both sets of rates, minor misdemeanors were excluded, including public intoxication, driving and traffic violations, gambling, vagrancy, and suspicion. In addition, rates were based on number of arrests, not on number of persons arrested. For both reasons, these rates are not directly comparable to those of other studies. Since only five female patients were arrested during the study period, analyses of results were based on males only.

Analysis of arrest by crime category is of doubtful reliability in view of the few cases in each of the 17 categories listed. Perhaps more meaningful is the pattern that emerges across categories: Patient arrest rates were higher in all categories of violent crime (crime against persons), four out of five types of crime against property, and drug offenses. They were lower for "white collar" crimes such as forgery, counterfeiting, and embezzlement.

The distribution of diagnoses among males in this Wyoming sample is uncommon (diagnoses of females were not given). Nearly 70% had a primary or secondary diagnosis of alcoholism, 14% were schizophrenic, 10% were personality disorders, and 3.5% were drug abusers. Surprisingly, in light of others' findings of an association between alcohol and crime, alcoholics were not overrepresented among arrested patients, nor were schizophrenics. Further, in this sample schizophrenics committed no violent crimes, in contrast to Zitrin et al.'s (1976) findings. The two groups with excessive arrest rates were patients with personality disorders and drug abusers.

Demographic variables associated with patient arrests included sex, age, marital status, and type of hospitalization. Single men aged 18 to 24 had a disproportionate share of arrests, as did those who were involuntarily hospitalized. Women constituted 38% of the patient group, but accounted for only 5% of the arrests.

Comparing arrest rates before and after hospitalization, two thirds of the arrests occurred before. Of the 49 men arrested during the 10-year period, 6 had no prior arrests and 35 had no subsequent arrests. In other words, the high recidivism noted elsewhere is not apparent here. Durbin et al. (1977) also noted a clustering of arrests in the year preceding hospitalization, suggesting that the disruptive behavior leading to the arrests may also have led to the decision to hospitalize. Here, as elsewhere, it is observed that hospitalization "may often serve as a diversionary adjunct to the criminal justice system" (p. 83).

The most recent investigation of arrest rates of discharged patients was conducted at the request of the New York State Commissioner of Mental Hygiene. Steadman, Melick, and Cocozza (Note 5) designed their study both to serve as a follow-up to Brill and Malz-

berg's (1962) 1947 study and to compare arrests of patients discharged in 1968 and 1975, before and after introduction of more liberal and rapid discharge policies in the state hospital system.

The 1968 sample consisted of every 14th patient aged 18 or more who was discharged from state facilities in that year, producing a study group of 1,920. In 1975, every 18th patient was selected, creating a sample of 1,938. In both samples, a small majority was male, 70% were white, 22% were black, and 7% were Puerto Rican. Most had not completed high school. Half in 1968 and 43% in 1975 lived in New York City. The most common diagnostic label was schizophrenia (27% in 1968 and 32% in 1975), followed by paranoid states (20% in both samples) and substance abuse (16% in 1968 and 19% in 1975). Over 80% of all the patients were released within 6 months of the current hospital admission. Nineteen percent of the 1968 sample and 26% of the 1975 sample had a history of prior arrests.

Lifetime histories of psychiatric hospitalization and criminal records before the current hospitalization were obtained from state agencies. The median follow-up period for both groups was 19 months (the briefest of any study reviewed). Patient arrest rates were compared with those of the general population of New York State over age 17.

Patients' total annual arrest rates were found to be more than double those of the general population in both years studied. In 1968 the patient rate per 1,000 was 73.5, compared to 27.5 for the general population. In 1975 the discrepancy was even greater: 98.5 for patients and 32.5 for the general population. Rates for both increased between occasions, but the patient rate accelerated more rapidly.

Analysis of arrests by crime category showed patient rates exceeding public rates within every category in both years with one exception: approximately equal rates for sex crimes in 1968. Greatest excess occurred in arrest rates for property, sex, and violent crimes in the 1975 sample and in arrest rates for drug, property, and violent crimes in the 1968 sample. During the 19 months after discharge, .9% of the 1968 sample and 1.7% of the 1975 sample committed violent crimes (homicides and assaults)—a small but increasing proportion.

The authors analyzed demographic, psychiatric, and historical factors associated with postdischarge arrests, first singly and then simultaneously. Taken one at a time, several variables were found to be significantly correlated with arrests, including youth, shorter hospital stays (presumably because the disorders were less serious or more transient), sex (fewer females arrested), ethnicity (in 1968 both blacks and Puerto Ricans had disproportionately high rates; in 1975 only blacks did), and admitting diagnosis. Personality disorders and substance abusers had the highest arrest rates. Although schizophrenics constituted about 25% of the sample in 1968 and 32% in 1975, only 3% and 6% were arrested in the respective years.

When multivariate analysis was applied to these data, diagnosis dropped out as a meaningful predictor; most of its relationship to arrest rates was accounted for by the variables of age and sex: Most personality disorders and substance abusers are young males. Using multiple regression analysis with the three variables with the highest zero order correlations to arrest rates—total prior arrests, age, and admitting diagnosis—Steadman et al. (Note 5) found that only the first two made independent contributions, together accounting for 13% of the variance in subsequent arrest behavior. Interaction effects were not analyzed as such.

In this study, as in virtually every other that looked at such data, a history of arrest before hospitalization turned out to be the best single predictor of postdischarge arrests so far identified. Here, as elsewhere, when arrest records of patients with and without prior arrests were separately analyzed and compared with those of the general population, postdischarge arrests of those with prior records substantially exceeded general population rates, whereas patients with no such history had lower rates than the general population in five out of the six crime categories studied (the exception was property crimes). Furthermore, Steadman and his

colleagues found a gradient such that patients with three or more prior arrests had more postdischarge arrests than those with two arrests, who in turn had more postdischarge arrests than those with one arrest. Not only was prior arrest itself a useful predictor but a pattern of similar offenses was also observed. Thus, the best predictor of subsequent arrest for violent crime was arrest for one or more violent crimes preceding hospitalization.

The predictive utility of a prior arrest history is widely recognized in the criminal justice literature. FBI studies (Federal Bureau of Investigation, 1973) of recidivism among over 200,000 persons arrested for federal offenses in the period 1970 through 1972 showed that 65% had been arrested at least twice, with an average of four arrests during the preceding 5 years. Furthermore, of these repeat offenders, 44% were rearrested in states other than that where first arrested. For the repeat offenders in 1972 alone, 55% of prior arrests took place in another state. Frequency of prior arrests was negatively associated with age; offenders under age 20 were arrested on the average every 3 months, whereas those over age 40 were arrested less than every 2 years. Parenthetically, such high interstate mobility among repeat offenders suggests the possibility of significant underenumeration of arrests among mental patients in those studies that relied only on single-state arrest data.

In Steadman et al.'s study, the total arrest rates for all patients combined are inflated by the presence of patients with criminal records. More such persons are found among the patient population of New York State mental hospitals than among the general population, and it seems to be this excess that accounts for higher total patient arrest rates. In addition, the proportion of patients admitted with arrest records has increased strikingly over time, from Brill and Malzberg's (1962) finding of 15% for male patients in 1947 to 32% of male patients in 1968 (Steadman et al., Note 5). By 1975, 4 out of 10 male patients had police records prior to their hospitalization. Approximately one third of the patients with arrest histories at each time period were arrested after discharge, compared to 2%–4% of patients without prior records who were subsequently arrested. As Steadman and his colleagues concluded, "Prior criminality rather than mental illness appears to be the primary explanation for the increasing arrest rates" (p. 16) observed.

Evaluation of the Evidence

Many questions about criminal behavior of mental patients have been raised and partially answered by the studies reviewed. It is, however, difficult to evaluate the evidence generated by each study when they are considered consecutively. In this section, therefore, I summarize relevant findings to clarify how much has been learned and what further information is needed.

Do Mental Patients Currently Have Higher Arrest Rates Than Members of the General Population?

Today and over the past 20 years, mental patients discharged from public facilities *as a group* have total arrest rates for all crimes that equal or exceed public rates with which they have been compared. Arrest and conviction rates for the subcategory of violent crimes were found to exceed general population rates in every study in which they were measured. The certainty of these findings must be tempered by consideration of design limitations including nonequivalent comparison groups and lack of attention to variables such as age and social class.

Have These Rates Changed Over Time?

The arrest rates of paroled and discharged patients based on pre-1950 records consistently were found to be lower than those reported for the general population. There has been a pronounced relative as well as absolute increase in arrests of mental patients since then; that is, whereas arrest rates for both patients and the general public have increased, the rate of acceleration for patients has been much greater.

What Factors Have Contributed to This Upward Trend in Arrest Rates?

At least two developments may account for the observed changes, both of which are fundamentally related to the evolving social role of the mental hospital in our society. First, there is the change in hospital policies regarding involuntary admission and retention of patients and the likelihood and timing of their discharge. In the past, there were comparatively few limitations regarding grounds for involuntary civil commitment. In contrast, recent legal reforms have largely restricted such commitment to situations in which the prospective patient is perceived as potentially dangerous to himself or others (Cocozza & Steadman, Note 2). Although psychiatric predictions of dangerousness are admittedly inaccurate (Steadman, 1973; Stone, 1975), it does seem probable that these restricted criteria alter the nature of the patient population and enhance the probability that discharged patients may be at greater risk for subsequent dangerous behavior.

Duration of commitment and discharge policies have also changed. Forty years ago, patients were often committed for decades or for life in large state facilities, physically removed from their communities of origin. As Pollock (1938) described in detail, patients who were released had to meet a series of requirements, not only with respect to their psychiatric status and social functioning but also regarding available familial resources. Once these recovered, socially acceptable, and cooperative patients with welcoming and protective friends and relatives were paroled, their progress was monitored by an elaborate, extensive, and compulsory system of home and clinic visits. If parole authorities were notified of a change in status, the patient could be rehospitalized at once. Although New York State's hospital system may have been more cautious about discharges than others (Giovannoni & Gurel, 1967), this general system prevailed nationally.

Today in New York State facilities as elsewhere, the average patient stay is 30 days. In acute-care units, patients are usually discharged within 2 weeks. Nationally since the 1950s hospital stays have become progressively shorter, and hospital populations have declined although admissions have not, leading to many more discharged patients in the community. Since the late 1960s heavy emphasis has been placed on community-based treatment on a voluntary basis. As soon as the acute symptoms of hospitalized patients subside, usually with the help of psychotropic medication, patients are discharged. There is no compulsory aftercare or follow-up. In short, virtually all patients admitted in the past several years are promptly returned to the community. The state mental hospital's traditional role as a "warehouse for the unwanted" has been transformed into a brief way station for the most acutely disturbed.

The second major change concerns a shift in the way the civil machinery distributes disruptive members of the community. In the past, a large majority of offenders were sent to jail regardless of their mental status. Since World War II it has become increasingly common for offenders who appear to be mentally disturbed or who have a history of psychiatric hospitalization to be dispatched to the mental health rather than to the criminal justice system. This phenomenon seems related to the fact that, in most jurisdictions, the criminal justice system is inundated with far more defendants than can be handled. Such overcrowding evidently has encouraged the use of psychiatric hospitalization as an alternative method of removing disturbing people from the community. Consequently, an increasing number of mental patients have police records. It is reasonable to expect for this reason alone that, as all patients are promptly discharged, mental patients returning to the community will, as a group, have a higher arrest rate than in the past when the patient population had a different composition. As Giovannoni and Gurel (1967) have stated the point, "It is primarily in the way mental hospitals are utilized by the community, and particularly as this influences the kinds of patients admitted and the number and kinds of patients released, that one is likely to find the major sources of variation in ex-patient crime rate" (p. 151).

What Are the Best Predictors of Postdischarge Arrests?

It has been repeatedly and convincingly demonstrated that the small subset of patients who have prior criminal records accounts for a large majority of postdischarge arrests. This is the single best predictor so far identified. Other predictors, which are significantly associated with postdischarge arrests but not large in their effect, include male sex, youth, and unmarried status. Short hospital stay and diagnoses of alcoholism, drug abuse, and personality disorder are also associated with arrest risk when analyzed separately, but their predictive value seems to be largely accounted for by their relationship to sex and age, as Steadman et al. (Note 5) demonstrated in their multivariate analyses. Despite the consistency with which these predictors have been identified, their effect size is evidently small. Much remains to be learned in the prediction of disruptive, illegal behavior by mental patients after their return to the community.

What is the Association Between Arrest Risk and Diagnostic Category?

It is generally believed that patients diagnosed as personality disorders, alcoholics, and drug abusers are the most likely to display antisocial and aggressive behavior. There is less consensus about whether these three diagnostic labels describe mental illness. In some epidemiological studies of true prevalence rates, investigators systematically exclude these categories (e.g., Gove & Tudor, 1973) and focus exclusively on neuroses and functional and organic psychoses. In addition, states vary in their policies of admitting patients with a primary diagnosis of alcoholism more than once to state facilities, so that their proportion of the total patient population varies by area as a function of administrative policy.

In those studies that generated interpretable findings about crime rates of different diagnostic groups in relation to their size in the patient sample under study, alcoholics, addicts, and personality disorders were in each case found to have excess rates (Brill & Malzberg, 1962; Durbin et al., 1977; Zitrin et al., 1976; Steadman et al., Note 5). The consistency of these results lends them credibility. The effects of age and sex must be considered, however, in view of Steadman et al.'s observations regarding their associations. Because two of the four studies dealing with diagnostic differences were based on male samples only (Brill & Malzberg, 1962; Durbin et al., 1977), sex is evidently not a critical factor. The role of age remains unclear. Giovannoni and Gurel (1967) found, in their group of older, chronic schizophrenic males, that nearly three quarters of those arrested had drinking problems. More direct evidence would be available by studying differential diagnostic roles in an age-stratified sample to see if the differences prevail at various ages. At present, one may conclude that patients with diagnoses of personality disorders, alcoholism, and drug dependency have disproportionate arrest rates. Whether this is caused by the nature of their disorders or by demographic characteristics associated with their distribution in the general population remains unclear.

Evidence is less consistent regarding the subsequent criminal activity of patients diagnosed as schizophrenic. In those studies in which schizophrenics constituted half or less of the patient sample, their arrest rates were not disproportionately high. In the most detailed analysis of arrests by diagnostic group, conducted by Zitrin et al., schizophrenics were found overrepresented among those patients who committed violent crimes with bodily harm but not overrepresented in terms of overall arrest rates for all crime categories combined. In two samples in which all or most of the patients were schizophrenic (Giovannoni & Gurel, 1967; Sosowsky, Note 3, San Mateo sample), arrest rates were much higher for patients than for control populations, especially for violent crimes; that is, schizophrenic samples had the same excess of crimes, compared to control groups, that characterized samples with predominantly nonschizophrenic patients. From the preliminary evidence available, it seems that arrest rates of schizophrenics do not exceed those of other diagnostic groups, with the possible exception of arrests for violent crimes (homi-

cides and assaults) in which their rates may be higher. Critical and uncontrolled are the factors of age and social class, which require evaluation before firmer conclusions can be derived.

Are Mental Patients More Likely To Be Arrested for Certain Types of Crime?

Traditionally, it was believed that mental patients, when arrested at all, were charged with minor offenses such as loitering, vagrancy, or public intoxication. It was an unpleasant surprise to learn not only that patient arrest rates were the same or greater than those of the general population in recent years but that this excess was particularly pronounced in the category of felonies, and specifically, of violent crimes or crimes against persons. Six investigators, using eight patient samples, found higher arrest (and/or conviction) rates for the crimes of homicide and assault among patients than among control groups (see Table 2). The magnitude of this excess ranged from 1½ to 29 times greater than general population rates. No studies reported contrary findings. The consistency and size of the differences across patient samples that vary in time, place, and diagnostic composition make the differences convincing. It seems reasonable to conclude that mental patients are more likely to be arrested for assaultive and sometimes lethal behavior than are other people.

No clear pattern emerges regarding the relative frequency with which other crimes are charged against mental patients. Rappeport and Lassen (1965, 1966) found the highest arrest rates for robbery by males and for aggravated assault by females. Giovannoni and Gurel (1967) reported that after violent crimes, which accounted for 27% of police contacts among their patients, the next most common offenses were drunkenness (23%), motor vehicle thefts (13%), and crimes against property (12%). Durbin et al. (1977) reported a high incidence of drug offenses among their predominantly alcoholic patients, and Steadman et al. (Note 5) found the most common offenses in their 1968 sample to be drug, property, and violent crimes; the most common offenses in their 1975 sample were property, sex, and violent crimes. In short, apart from assaultive behavior, mental patients are apparently no more likely to be arrested for some crimes than others, judging from the limited evidence at hand.

Does Hospitalization Reduce the Probability of Recidivism Among Patients With Prior Arrest Records?

This question is interesting because a positive finding would provide support for the notion of a causal association between mental status and crime, in contrast to the opinion of several investigators that disturbed and criminal behaviors may coexist but are causally independent. No definite answers can be generated by analysis of recidivism rates among patients in the absence of data regarding recidivism among nonhospitalized offenders. The only authors who tried to assess the latter were Brill and Malzberg (1962), but they did not have access to reliable recidivism data for the general criminal population.

Despite the absence of control data, several authors have discussed the issue of recidivism rates among patients. Cohen and Freeman (1945) were the first to suggest that hospitalization may reduce recidivism, based on their finding that only 26% of patients arrested before being hospitalized were also arrested after discharge. The only others to be impressed with a decline in arrests after hospitalization among patients with records were Durbin and his colleagues (1977). Of the 43 men in their sample who were arrested in the 5 years before hospitalization, only 8, or 19%, were also arrested within 5 years after discharge. They concluded that "factors associated with hospitalization . . . may have influenced the reduction in arrest rates after hospitalization" (p. 83).

Other investigators who discussed recidivism disagreed. Rappeport and Lassen (1965) stated, "As best we can interpret our data, there was a tendency toward recidivism not unlike that seen in the general community." Brill and Malzberg (1962) and Steadman et al. (Note 5) interpreted their findings similarly. Zitrin et al. (1976) only reported

Table 6
Postdischarge Arrest Rates of Mental Patients With and Without Prior Arrest Records

Authors	Years of discharge	No. of patients with prior records	Follow-up period	Percentage arrested with prior records	Percentage arrested without prior records
Cohen & Freeman (1945)	1940 to 1944	314 (18% of sample)	$M = 2$ years	26%	<1%
Brill & Malzberg (1962)	1947	803 males	5½ years	34%	2%
Zitrin, Hardesty, Burcock, & Drossman 1976)	1969 to 1971	64 (only those arrested for violent crimes)	2 years	56%	—
Steadman, Melick, & Cocozza (Note 5)	1968	343 males	19 months	31%	2%
Steadman, Melick, & Cocozza (Note 5)	1975	435 males	19 months	29%	4%
Durbin, Pasewark, & Albers (1977)	1969	43	5 years	19%	3%

recidivism rates for the 85 patients in their sample arrested for crimes of violence for 2 years before or after the hospitalization under study. One third were arrested only before, one quarter only after, and 42% both before and after their hospital stay, producing a total recidivism rate of 56% among patients who committed one or more crimes of violence. This rate is the highest of any reported, as seen in Table 6.

In summary, reported recidivism rates for arrests among discharged mental patients range from 19% to 56%. These rates were derived from different periods of record surveillance, include different crime categories, and were gathered over a period of 30 years. In contrast, only 2%–4% of patients without prior arrest records were arrested within 5 years after discharge. This analysis of recidivism rates reinforces the conclusion noted earlier that a history of prior arrests is a useful predictor of arrest risk after discharge. Without equivalent recidivism rates for the nonhospitalized criminal population, no conclusions are warranted regarding the impact of hospitalization on the risk of future antisocial behavior.

Conclusions

From the information presently available, it seems that discharged mental patients as a group are not significantly less likely than others to exhibit dangerous or illegal behavior. At the present time there is no evidence that their mental status as such raises their arrest risk; rather, antisocial behavior and mentally ill behavior apparently coexist, particularly among young, unmarried, unskilled, poor males, especially those belonging to ethnic minorities. It is unlikely that most people would care to have such neighbors even in the absence of a history of psychiatric hospitalization.

The major factor associated with increases in arrest rates of discharged mental patients in recent years is the increased proportion of mental patients who have arrest histories before their hospitalization. For males in New York State, for example, this proportion

has risen from 15% to 40% in the past 30 years. This change of hospital clientele seems to represent an alteration in application of the civic machinery according to which the mental health system is being increasingly used as an adjunct to the criminal justice system.

Arrests are fairly infrequent events even when mental patients are inappropriately considered as a single group. In the 18 months after discharge from New York State hospitals in 1975, for example, 90% of patients were not arrested. When patients with arrest histories, primary diagnoses of substance abuse, and personality disorders are considered separately, the remainder of the patient group appears to be considerably less dangerous than are those members of the general public who are not mentally ill.

It is recommended that future studies emphasize selection of equivalent comparison groups, integration of patient data with arrest and recidivism rates of the nonhospitalized population, greater consideration of geographic issues, and improved data analysis. The majority of both mental patients in public institutions and criminals come from similar population subgroups. Both have disproportionate numbers of poor, unskilled, uneducated, unmarried young men, many of whom belong to low-status ethnic groups. If arrest rates of mental patients were compared to those of their nonhospitalized peers using these criteria, it is quite possible that the observed excess of arrests of discharged mental patients would no longer be apparent. This does not detract from the validity of the conclusion that as a group mental patients are more likely to be arrested, especially for crimes of violence, than is the total population, the majority of whom are obviously not poor, unmarried males that belong to ethnic minorities.

Based on the limited evidence available, I conclude that patients discharged from mental hospitals are not, by virtue of their psychiatric disorders or hospitalization experience, more prone to engage in criminal activity than are people demographically similar to them who do not have a history of mental illness. Although patients considered as a group do have higher arrest rates than nonpatients considered as a group, it is largely because the patients include in their midst a disproportionate share of people with prior police records. The most immediately obvious method of reducing criminal activity among discharged mental patients is to reexamine and modify current procedures that contribute inappropriately to the use of mental hospitals as alternatives to the criminal justice system.

Reference Notes

1. Piasecki, J. *Community response to residential services for the psycho-socially disabled: Preliminary results of a national survey.* Paper presented at the First Annual Conference of the International Association of Psycho-Social Rehabilitation Services, Philadelphia, November 1975.
2. Cocozza, J., & Steadman, H. *Community fear of the mentally ill: An unsolved obstacle for the community mental health movement.* Paper presented at the Conference on Community and Policy Research, Albany, New York, April 1976.
3. Sosowsky, L. Violence and the mentally ill. In, *Putting state mental hospitals out of business—The community approach to treating mental illness in San Mateo County.* Berkeley: Graduate School of Public Policy, University of California, July 1974.
4. Zitrin, A. Personal communication, July 1978.
5. Steadman, H., Melick, M., & Cocozza, J. *Arrest rates of persons released from New York State Department of Mental Health Hygiene psychiatric centers.* Report to the New York State Commissioner of Mental Hygiene, 1977.

References

Ashley, M. Outcome of 1000 cases paroled from the Middletown State Hospital. *State Hospital Quarterly,* 1922, *8,* 64–70.

Aviram, U., & Segal, S. Exclusion of the mentally ill: Reflection of an old problem in a new context. *Archives of General Psychiatry,* 1973, *29,* 126–131.

Brill, H., & Malzberg, B. Statistical report on the arrest record of male ex-patients released from New York State mental hospitals during the period 1946–8. In, *Criminal acts of ex-mental hospital patients* (Supplement No. 153). Washington, D.C.: American Psychiatric Association Mental Hospital Service, August 1962.

Cohen, L., & Freeman, H. How dangerous to the community are state hospital patients? *Connecticut State Medical Journal,* 1945, *9,* 697–700.

Cumming, J., & Cumming, E. On the stigma of mental illness. *Community Mental Health Journal,* 1965, *1,* 135–143.

Durbin, J., Pasewark, R., & Albers, D. Criminality and mental illness: A study of arrest rates in a rural state. *American Journal of Psychiatry,* 1977, *134,* 80–83.

Farnsworth, D. Dangerousness. *Psychiatric Annals,* 1977, *7,* 55–70.

Federal Bureau of Investigation. *Uniform crime reports for the United States, 1972.* Washington, D.C.: U.S. Government Printing Office, 1973.

Federal Bureau of Investigation. *Uniform crime reports for the United States, 1976.* Washington, D.C.: U.S. Government Printing Office, 1977.

Giovannoni, J., & Gurel, L. Socially disruptive behavior of ex-mental patients. *Archives of General Psychiatry,* 1967, *17,* 146–153.

Gove, W. R., & Tudor, J. Adult sex roles and mental illness. *American Journal of Sociology,* 1973, *78,* 812–835.

Gulevich, G., & Bourne, P. Mental illness and violence. In D. Daniels, G. Mapshall, & F. Ochberg (Eds.), *Violence and the struggle for existence.* Boston: Little, Brown, 1970.

Lagos, J., Perlmutter, K., & Saexinger, H. Fear of the mentally ill: Empirical support for the common man. *American Journal of Psychiatry,* 1977, *134,* 1134–1137.

Levine, D. Criminal behavior and mental institutionalization. *Journal of Clinical Psychology,* 1970, *26,* 279–284.

Lilienfeld, A. *Foundations of epidemiology.* New York: Oxford University Press, 1976.

Nunnally, J. *Popular conceptions of mental health: Their development and change.* New York: Holt, Rinehart & Winston, 1961.

Paull, D., & Malek, R. Psychiatric disorders and criminality. *Journal of the American Medical Association,* 1974, *228,* 1369.

Pollock, H. H. Is the paroled patient a threat to the community? *Psychiatric Quarterly,* 1938, *12,* 236–244.

Rabkin, J. Public attitudes toward mental illness: A review of the literature. *Schizophrenia Bulletin,* 1974, No. 10, 9–33.

Rabkin, J. The role of attitudes toward mental illness in evaluation of mental health programs. In M. Guttentag & E. Struening (Eds.), *Handbook of evaluation research* (Vol. 2). Beverly Hills, Calif.: Sage, 1976.

Rappeport, J., & Lassen, G. Dangerousness-arrest rate comparisons of discharged patients and the general population. *American Journal of Psychiatry,* 1965, *121,* 776–783.

Rappeport, J., & Lassen, G. The dangerousness of female patients: A comparison of the arrest rate of discharged psychiatric patients and the general population. *American Journal of Psychiatry,* 1966, *123,* 413–419.

Rappeport, J., Lassen, G., & Hay, N. A review of the literature on the dangerousness of the mentally ill. In J. R. Rappeport (Ed.), *Clinical evaluation of the dangerousness of the mentally ill.* Springfield, Ill.: Charles C Thomas, 1967.

Segal, S., & Aviram, U. Community-based sheltered care. In P. Ahmed & S. Plog (Eds.), *State mental hospitals.* New York: Plenum Press, 1976.

Sosowsky, L. Crime and violence among mental patients reconsidered in view of the new legal relationship between the state and the mentally ill. *American Journal of Psychiatry,* 1978, *135,* 33–42.

Steadman, H. Some evidence on the inadequacy of the concept and determination of dangerousness in law and psychiatry. *Journal of Psychiatry & Law,* 1973, *1,* 409–426.

Stone, A. *American Journal of Psychiatry,* 1975, *132,* 829–831.

Zitrin, A., Hardesty, A., Burdock, E., & Drossman, A. Crime and violence among mental patients. *American Journal of Psychiatry,* 1976, *133,* 142–149.

Received October 11, 1977 ■

Call for Nominations

The APA Publications and Communications Board invites nominations for the editorship of *Psychological Bulletin* for the years 1981 through 1986. R. J. Herrnstein is the incumbent editor. The new editorial appointment will be made this year in order to provide continuity in publication of the journal. The editor-elect will start to receive manuscripts in 1980 to prepare for issues published in 1981, the first year of the new editorial term.

To nominate candidates, prepare a brief statement of one page or less in support of each nomination. Submit nominations no later than April 1, 1979, to Anita DeVivo, P&C Board Liaison, APA, 1200 Seventeenth Street, N.W., Washington, D.C. 20036.

Psychological Bulletin
1979, Vol. 86, No. 1, 28–32

Comment on Banks's "White Preference in Blacks: A Paradigm in Search of a Phenomenon"

John E. Williams
Wake Forest University

J. Kenneth Morland
Randolph-Macon Woman's College

Banks concluded from the studies he reviewed that the evaluative preference and self-identification responses of Afro-American children toward stimulus alternatives representing light- and dark-skinned persons conformed to simple chance rather than indicating a "white preference in blacks." This interpretation is challenged as misleading because of Banks's dismissal of the importance of comparisons by race in the literature cited and because of his failure to cite a number of relevant studies of race and color bias, the results of which are inconsistent with his conclusion.

In a recent *Psychological Bulletin* article, Banks (1976) reviewed a number of studies in which young Afro-American [1] children had made preference and self-identification responses to light- and dark-skinned persons, as represented by dolls, puppets, line drawings, and photographs. Banks concluded that the findings of these studies, viewed in toto, could be attributed to chance and, hence, provided no systematic evidence of a bias favoring light-skinned persons among Afro-American children. Having been active in this area of research, we feel that Banks's conclusion is misleading and requires a reply. We are concerned primarily with two matters: first, Banks's abrupt dismissal of the relevance of comparing the responses of Afro-American children to those of Euro-American children and other groups, with a consequent distortion of the significance of much of the research cited; second, Banks's omission of the findings of additional relevant studies that are in conflict with his conclusion, particularly recent and more methodologically sophisticated studies of preschool children's attitudes toward race and color.

We begin by acknowledging that Banks was correct in pointing out that in most of the studies he cited the racial preference and self-identification responses of Afro-American children can be attributed statistically to chance. This was the case even when a majority of the Afro-American children gave "white-preference" responses, as for example in the picture–interview studies cited in which more than half of preschool Afro-American respondents indicated that they would prefer to play with the Euro- rather than with the Afro-American models, that they looked more like the Euro- than the Afro-American models, and that their mothers looked more like the Euro- than the Afro-American models (Morland, 1962, 1963). This was also true for another study Banks cited in which preschool Afro-Americans who had demonstrated they could use racial classification terms [2] correctly were asked to which race they themselves belonged: 52% responded correctly, 16% did not know or refused to say, and 32% responded incorrectly (Morland, 1958). In other words, half of the Afro-American

Requests for reprints should be sent to John E. Williams, Department of Psychology, Wake Forest University, Winston-Salem, North Carolina 27109.

[1] We prefer to use the terms *Afro-American* and *Euro-American* rather than *black* and *white* to designate these racial categories for reasons we have explained elsewhere (Williams & Morland, 1976, pp. x–xi).

[2] In such studies, the racial classification terms best known to the children are checked empirically at the time of each testing. The racial designation of Euro-Americans as *white* has not changed; however, racial designations of Afro-Americans have changed. In the 1958 study it was *colored;* later it shifted to *Negro;* now it is *black*.

children made correct self-classification by race, and half did not. Subsequent studies of preschool Afro-American children have yielded similar results (Morland, 1963, p. 239; 1969, p. 368; Savory, Note 1).

One might assume from such apparently inconsistent responses that racial self-classification is unimportant in American society and/or that preschool American children are not old enough to recognize and attach significance to racial differences. There is considerable evidence, however, to show that racial classification has been and continues to be important in American society, and, further, the assumption can be supported that a societal norm exists that calls for Americans to identify with and be proud of the race to which society says they belong (Williams & Morland, 1976, pp. 3–32, 251). Are preschool American children old enough to have learned to recognize racial differences and to know their own racial classification? It is at this point that the comparison by race becomes important. In the Morland studies referred to above, Euro-American preschoolers did not respond randomly; rather, they chose the Euro- over the Afro-American models in every response at statistically significant levels, according to the criterion percentages employed by Banks (1976, Table 1, p. 1180). The Euro-American preschoolers chose the Euro- rather than the Afro-American models as the ones they preferred to play with, the ones they looked more like, the ones they would rather be, and the ones their mothers looked more like; and of those who had demonstrated that they knew how to use racial classification terms, 99.5% said they were "white" (Morland, 1958, 1962, 1963). It is clear from these findings that preschool American children can learn to recognize race differences and make correct racial self-classification.

While Banks (1976) realized that differences in response by race had been found, he appears to have dismissed the relevance of comparisons by race because of his disagreement with what he saw as their implicit rationale, namely that "same-race choices of white subjects may be believed to represent an a priori standard of rational behavior" or "a standard of mental health" (p. 1180). There is, of course, a more general rationale for making these comparisons by race. It is the desire to increase understanding of the development of racial preference and identity by seeing if they are significantly related to the race of the respondent. The same rationale applies to comparisons by age, sex, socioeconomic status, and region, which are found in the studies Banks cites. If it is granted that racial classification is important in American society, that a societal norm for own-race identification exists, and that preschool American children are old enough to recognize and attach significance to racial differences, the question arises from the studies Banks reviewed as to what else there is about American society and the nature of racial bias that leads to such different responses in young American children by race. Banks avoided dealing with this question, which we believe deserves serious pursuit. We have done this by comparing racial identity and attitude responses of children in other societies, noting how social structure, norms, and reactions to color are related to racial self-identity and attitudes (Morland, 1969; Morland & Williams, 1969; Williams, Morland, & Underwood, 1970; Morland, Note 2). From these and other studies, we have constructed what we hope will be a useful theoretical model for increasing understanding of the development of race bias in young children (Williams & Morland, 1976, pp. 280–283).

Related to Banks's (1976) dismissal of comparison by race is the inaccuracy that comes in citing the 1958 and 1963 studies by Morland to support his assertion: "Reliance upon white comparative frames has largely perpetuated the notion of black self-rejection . . ." (p. 1185). In neither of these studies was any mention made of racial self-rejection, and it is indeed strange that Banks failed at this point to bring in the 1962 study by Morland, which Banks used in other places in his article. That study explicitly stated as one of its conclusions: "Preference for one race did not imply rejection of the other" (Morland, 1962, p. 279). This conclusion was based on the racial acceptance measure in which no choice was involved. Both the Afro- and Euro-

American respondents readily accepted models of both races, and rejection based on race was exceedingly rare. This absence of racial self-rejection has also been explicitly pointed out in other studies by Morland (1966, p. 26; 1969, pp. 364–366; Note 3, pp. 5–6; Note 4, p. 6) and in the review by Williams and Morland (1976, pp. 191–192).

In analyzing the studies he reviewed, Banks did not make a distinction between responses of preschool and in-school Afro-Americans. We have discovered that unless this is done, the conclusions can be misleading. In our analysis of a number of studies, including several of those cited by Banks, we found that in-school Afro-American children were significantly more likely than preschool Afro-Americans to prefer and identify with Afro-American models (Williams & Morland, 1976, pp. 176–201). For example, less than 2% of in-school Afro-American children of high racial classification ability gave incorrect racial self-classification responses. This was in marked contrast to the preschool Afro-Americans of high racial classification ability of whom as many as 30% gave incorrect racial self-classification responses. This difference between preschool and in-school Afro-American children questions Banks's (1976) contention that the high level of own-race preference and identification by Euro-American children could be an expression of their "ethnocentrism" (pp. 1180, 1185). By the time Afro-American children enter public school, they evidently respond in a similar way to Euro-American children in own-race preference and identification. Rather than being "ethnocentric," these responses can be accounted for as a recognition by the children of the racial category to which American society says they belong and as an expression of the American norm that persons should be proud of and accept their racial affiliation.

Banks's (1976) paper was submitted in 1975 and the studies he reviewed appear to cover the years from 1939 through 1973 (Table 3, p. 1184). We have already cited several studies of that period that were not referred to, for example, those by Morland in 1966 and 1969. Banks also omitted a number of research studies of racial attitudes in Afro-American children published during 1973–1975, studies that used different techniques from those he cited: Best, Smith, Graves, and Williams, 1975; Mabe and Williams, 1975; Spencer and Horowitz, 1973; Williams, Best, and Boswell, 1975; Williams, Best, Boswell, Mattson, and Graves, 1975; and Williams, Williams, and Beck, 1973. Also neglected were several doctoral dissertations and master's theses: Baugher, 1973; H. P. McAdoo, 1970; J. L. McAdoo, 1970; Skinto, 1969; Vocke, 1971; Walker, 1971; Whiteside, 1975. We have reviewed the findings from the foregoing studies elsewhere (Williams & Morland, 1976). Without exception, these studies have provided evidence of a tendency for young Afro-American children to evaluate representations of light-skinned persons more favorably than those of dark-skinned persons.

Three of these studies will be noted for illustrative purposes. H. P. McAdoo (1970) administered the 12-item racial attitude procedure devised by Williams and Roberson (1967) to preschool Afro-American children in Michigan and Mississippi. It was found that the mean scores in both groups departed significantly from chance in the direction of pro-light-skin bias. A second study was conducted by Spencer and Horowitz (1973), who studied the modification of color and race bias among Afro- and Euro-American preschoolers in Kansas. Although the modification procedures were found to be generally effective, the authors reported that both Afro- and Euro-American children in the nontreated control group displayed a bias for light-skinned persons throughout the study. A third study is that by Williams, Best, Boswell, Mattson, and Graves (1975), who described the development and standardization of the Preschool Racial Attitude Measure II (PRAM II). In this 24-item picture–story procedure, the child selects between drawings of light- and dark-skinned persons as the one described in an accompanying story that contains 1 of 24 evaluative adjectives. The PRAM II scores have a 24-point range and, thus, provide a more sensitive measure of attitude than the "one-item tests" reviewed by Banks (1976, Table 2, p. 1182). When PRAM II was administered to Afro-American preschoolers in North Carolina, the mean scores were found to depart significantly from chance in the direction of pro-light-skin bias but not to as ex-

treme a degree as among Euro-American children of comparable age.

Additional light has been thrown on the development of race bias in children through research on racial attitudes in other societies. As illustrations, we note the recent studies employing the PRAM II procedure with preschool children in France, Italy, Germany, and Japan, who have had little direct contact with dark-skinned persons or with concepts of race associated with color (Best, Field, & Williams, 1976; Best, Naylor, & Williams, 1975; Iwawaki, Sonoo, Williams, & Best, 1978). The findings from these studies indicated that these foreign children also displayed pro-light-skin bias, which suggests to us that contact with racial concepts of the type encountered in the United States is not necessary for the development of such bias.

Banks (1976) stated in the abstract of his article that his concern was with the "choice behavior among blacks toward white and black stimulus alternatives" (p. 1179). In view of this, it is surprising that he made no reference to studies of the evaluative responses of Afro-American children to the colors black and white (Skinto, 1969; Stabler, Johnson, Berke, & Baker, 1969; Stabler, Johnson, & Jordan, 1971; Vocke, 1971; Williams, Boswell, & Best, 1975; Williams & Rousseau, 1971). As an illustration of the findings from these studies, Williams and Rousseau (1971) administered a 12-item picture–story procedure to Afro-American preschoolers who were asked to choose between a white or a black animal as the one described in the associated story. It was found that the children displayed a significant tendency toward the association of positive evaluative adjectives with the color white and negative evaluative adjectives with the color black. Similar findings were reported by Williams, Boswell, and Best (1975), although the degree of prowhite bias among Afro-American children was not as extreme as that found among Euro-American children of comparable age. We think it is instructive also to note that a similar prowhite bias has been documented among preschool-aged children in a number of other countries including England, Scotland, France, Germany, Italy, and Japan (Best, Field, & Williams, 1976; Best, Naylor, & Williams, 1975; Dent, 1976; Iwawaki, Sonoo, Williams, & Best, 1978). These findings appear to parallel the findings from studies with young adults, which point to a pan-cultural tendency to evaluate white more positively than black (Adams & Osgood, 1973; Williams, Morland, & Underwood, 1970).

We agree with Banks that none of these findings should be interpreted as evidence of "racial self-rejection" by Afro-American children, a view that is supported by the research findings on race and self-esteem that have failed to indicate any consistent differences between Afro- and Euro-American children. Our theory, which is elaborated elsewhere (Williams & Morland, 1976), is that when pro-light-skin bias and pro-white-color bias are encountered in preschool children these phenomena can be viewed, most parsimoniously, as reflections of a general prolight/antidark bias. We propose that this bias originates in most young humans as a result of early learning experiences (e.g., with the light of day and dark of night) and is subsequently reinforced through social processes, occurring in almost all cultures, in which light is used to symbolize "goodness" and dark to symbolize "badness."

By way of summary, we feel that Banks's refusal to deal with comparisons by race in the studies he reviewed and his failure to consider the findings from other more methodologically sophisticated studies of attitudes toward race and color have led him to an over-simplified and misleading conclusion. Pro-light-skin bias among preschool Afro-Americans is not a particularly powerful phenomenon. It exists, however, and needs to be recognized by all persons who are concerned with the development of young children in our multiracial American society.

Reference Notes

1. Savory, Laina. *A test of the developmental theory of color and race bias.* Honors Thesis, 1976, Randolph-Macon Woman's College.
2. Morland, J. K. *The development of racial-ethnic awareness in Chinese and Americans* (Final Report, Grant No. OIP75-01184). Washington, D.C.: National Science Foundation, 1976.
3. Morland, J. K. *Racial attitudes in school children: From kindergarten through high school.* (Final Report, Project 2-C009). Washington, D.C.: U.S. Department of Health, Education, and Welfare, Office of Education, 1972.

4. Morland, J. K. *Racial attitudes and racial balance in public schools: A case study of Lynchburg, Virginia.* (Final Report, Grant No. NIE-G-76-0040). Washington, D. C.: U.S. Department of Health, Education, and Welfare, National Institute of Education, 1976.

References

Adams, F. M., & Osgood, C. E. A cross-cultural study of the affective meanings of color. *Journal of Cross-Cultural Psychology,* 1973, *4,* 135–156.

Banks, W. C. White preference in blacks: A paradigm in search of a phenomenon. *Psychological Bulletin,* 1976, *83,* 1179–1186.

Baugher, R., Jr. *The skin color gradient as a factor in the racial awareness and racial attitudes of preschool children.* Master's thesis, California State University, Fresno, 1973.

Best, D. L., Field, J. T., & Williams, J. E. Color bias in a sample of young German children. *Psychological Reports,* 1976, *38,* 1145–1146.

Best, D. L., Naylor, C. E., & Williams, J. E. Extension of color bias research to young French and Italian children. *Journal of Cross-Cultural Psychology,* 1975, *4,* 390–405.

Best, D. L., Smith, S. C., Graves, D. J., & Williams, J. E. The modification of racial bias in preschool children. *Journal of Experimental Child Psychology,* 1975, *20,* 193–205.

Dent, E. *Evaluative responses of preschool children to the colors black and white.* Unpublished master's thesis, University of Strathclyde, 1976.

Iwawaki, S., Sonoo, K., Williams, J. E., & Best, D. L. Color bias among young Japanese children. *Journal of Cross-Cultural Psychology,* 1978, *9,* 61–73.

Mabe, P. A., III, & Williams, J. E. Relation of racial attitudes to sociometric choices among second grade children. *Psychological Reports,* 1975, *37,* 547–554.

McAdoo, H. P. *Racial attitudes and self-concepts of black preschool children.* Unpublished doctoral dissertation, University of Michigan, 1970.

McAdoo, J. L. *An exploratory study of racial attitude change in black preschool children using differential treatments.* Unpublished doctoral dissertation, University of Michigan, 1970.

Morland, J. K. Racial recognition of nursery school children in Lynchburg, Virginia. *Social Forces,* 1958, *37,* 132–137.

Morland, J. K. Racial acceptance and preference of nursery school children in a southern city. *Merrill Palmer Quarterly of Behavior and Development,* 1962, *8,* 271–280.

Morland, J. K. Racial self-identification: A study of nursery school children. *The American Catholic Sociological Review,* 1963, *24,* 231–242.

Morland, J. K. A comparison of race awareness in Northern and Southern children. *American Journal of Orthopsychiatry,* 1966, *36,* 22–31.

Morland, J. K. Race awareness among American and Hong Kong Chinese children. *American Journal of Sociology,* 1969, *75,* 360–374.

Morland, J. K., & Williams, J. E. Cross-cultural measurement of racial and ethnic attitudes by the semantic differential. *Social Forces,* 1969, *48,* 107–112.

Skinto, S. M. *Racial awareness in Negro and Caucasian elementary school children.* Unpublished master's thesis, West Virginia University, 1969.

Spencer, M. B., & Horowitz, F. D. Effects of systematic social and token reinforcement on the modification of racial and color concept attitudes in black and white preschool children. *Developmental Psychology,* 1973, *9,* 246–254.

Stabler, J. R., Johnson, E. E., Berke, M. A., & Baker, R. B. The relationship between race and perception of racially related stimuli. *Child Development,* 1969, *40,* 1233–1239.

Stabler, J. R., Johnson, E. E., & Jordan, S. E. The measurement of children's self-concepts as related to racial membership. *Child Development,* 1971, *42,* 2094–2097.

Vocke, J. M. *Measuring racial attitudes in preschool Negro children.* Unpublished master's thesis, University of South Carolina, 1971.

Walker, P. *The effects of hearing selected children's stories that portray blacks in a favorable manner on the racial attitudes of groups of black and white kindergarten children.* Unpublished doctoral dissertation, University of Kentucky, 1971.

Whiteside, R. R., Jr. *The modification of black/white color attitudes and its effect upon racial attitudes as measured by the Preschool Racial Attitude Measure II.* Unpublished master's thesis, East Carolina University, 1975.

Williams, J. E., Best, D. L., & Boswell, D. A. Children's racial attitudes in the early school years. *Child Development,* 1975, *46,* 494–500.

Williams, J. E., Best, D. L., Boswell, D. A., Mattson, L. A., & Graves, D. J. Preschool Racial Attitude Measure II. *Educational and Psychological Measurement,* 1975, *35,* 3–18.

Williams, J. E., Boswell, D. A., & Best, D. L. Evaluative responses of preschool children to the colors white and black. *Child Development,* 1975, *46,* 501–508.

Williams, J. E., & Morland, J. K. *Race, color, and the young child.* Chapel Hill: University of North Carolina Press, 1976.

Williams, J. E., Morland, J. K., & Underwood, W. L. Connotations of color names in the United States, Europe, and Asia. *Journal of Social Psychology,* 1970, *82,* 3–14.

Williams, J. E., & Roberson, J. K. A method for assessing racial attitudes in preschool children. *Educational and Psychological Measurement,* 1967, *27,* 671–689.

Williams, J. E., & Rousseau, C. A. Evaluation and identification responses of Negro preschoolers to the colors black and white. *Perceptual and Motor Skills,* 1971, *33,* 50–54.

Williams, K. H., Williams, J. E., & Beck, R. C. Assessing children's racial attitudes via a signal detection model. *Perceptual and Motor Skills,* 1973, *36,* 587–598.

Received June 17, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 33–36

On the Importance of White Preference and the Comparative Difference of Blacks and Others: Reply to Williams and Morland

W. Curtis Banks
Social Learning Laboratory, Educational Testing Service
Princeton, New Jersey

Gregory V. McQuater and Jenise A. Ross
Princeton University

It has been contended that, in contrast to the argument set forth by Banks (1976), the phenomenon of white preference does obtain among blacks, albeit minimally, and further that such a phenomenon is important in its relationship to and implications for the social adaptation of such persons. The present paper argues, however, that even within those examples set forth of the significant demonstration of white preference in blacks, insufficient evidence is presented to reject the null hypothesis of nonpreference. It is further considered whether empirical evidence supports the validity of such a phenomenon as a measure of self-concept or social behavior preferences and whether therefore the "importance" of white preference as it obtains in blacks, or differentiates blacks from whites, can be sustained.

White preference among blacks is, as stated by Williams and Morland (1979), "not a particularly powerful phenomenon" (p. 31). A serious question is whether it is a phenomenon at all (see Banks, 1976), although it must be conceded that it can be expected, as much as any other event of behavior, at least to occur in some of the people some of the time. In this regard, Williams and Morland (1979) have presented certain evidence omitted from an earlier discussion by Banks, which they argue provides support both for the existence and the importance of the phenomenon of white preference in blacks. However, we should consider this evidence in light of the accepted criteria by which we separate those things that exist some of the time from those that have reliable and predictable occurrence in the behavior of a specifiable population and those things that are felt to be important from those that have a demonstrable validity. Even if it were considered for the moment that such a phenomenon does exist, the question of its meaning and importance would remain. Since in the analysis of importance, the question of existence must inevitably be raised, we focus here primarily upon the question of importance with respect to the three specific examples of that evidence cited by Williams and Morland (McAdoo, 1970; Spencer & Horowitz, 1973; Williams, Best, Boswell, Mattson, & Graves, 1975) as supportive of both the existence and the importance of white preference in blacks. For this purpose we might take, for example, the framework suggested by Cronbach and Meehl (1955) in which there are three levels at which we may ask the question of whether white preference, as measured within the common paradigms, is a valid phenomenon among blacks.[1]

The authors thank J. Rudy for his helpful comments on an earlier draft of this paper.

Requests for reprints should be sent to W. Curtis Banks, Social Learning Laboratory, Educational Testing Service, Princeton, New Jersey 08541.

[1] *Webster's New Collegiate Dictionary* (1974) defines the word "important" as pertaining to something that is "valuable in content or relationship," which is close to the concept of validity as it applies to the content or criterion significance of psychological measures.

Content Validity

The content validity of a measure of white preference would rely upon the specification of a universe of content that is accepted as defining the phenomenon and the demonstration that the measure constitutes a sample of that universe. Such measures as those in which subjects are asked to express choices of dolls, puppets, and hypothetical other children may, in this sense, derive their validity solely from the fact that they sample responses directly representative of those to which they wish to generalize in the real world. However, the overall content validity of such measures as indicative of white preference in blacks would rest upon both the representativeness of the stimulus items themselves and the ability of such items to evoke white-preference choices. One instrument whose ability in this regard has already been asserted is described by Williams, Best, Boswell, Mattson, and Graves (1975).

In their procedure, subjects are asked to make preferential selections between white- and black-representative stimuli.

Williams et al. (1975) reported a binomial test of probability for the within-subjects frequencies of white-stimulus choices across the 24 replicated items of their Preschool Racial Attitude Measure II (PRAM) instrument. Moreover, they reported that 60% of white subjects selected the white-representative stimuli more often than would be expected by chance across the 24 within-subjects replications, as did 39% of black children. However, their contention of the significance of the choice behavior of 39% of blacks who chose the white stimuli on 17 or more of the items, while intuitively compelling, is statistically inappropriate, due to the nonindependence (within-subjects derivation) of those frequency observations. While this within-subjects replication of preference choices may yield a more reliable categorization of respondents into white-preference and other classifications than did the various one-response procedures of past research, it is these between-subjects category frequencies that are appropriately testable by binomial and related analyses. In this regard, fully 72% of white children chose the white stimuli at (within-subjects) rates that led Williams et al. (1975) to label them as indicating "definite" or "probable" bias. While 52% of black subjects fell within these two categories, fully 58% could have been expected to do so by chance alone.

Williams and Morland have also referred to the "significant" white-preference frequencies reported by Spencer and Horowitz (1973). The choices of the control-group blacks to which Williams and Morland refer as having displayed white preference were reported only in terms of the mean percentage of white choices (approximately 70%). Within a sample of eight, this magnitude of frequencies across subjects would not reject the null hypothesis, since a percentage of roughly 85 would be required at the .05 level. However, since the data were reported as mean percent of choices within subjects, we might instead attempt to assess its significance by a t-test comparison against 50%. Using the error term from the analysis of variance reported by Spencer and Horowitz (1973), this test would yield a t value of approximately 1.31, hardly approaching significance even at the .05 level. Consequently, the phenomenon of white preference in blacks fails to obtain in this example as well. In fact, the instrument devised by Williams et al. (1975) may contain stimulus items fully representative of the domain to which we may wish to generalize, but the evidence derived from its application provides empirical support for a content-valid measure of nonpreference in blacks, quite contrary to the assertion of Williams and Morland.

Criterion Validity

At another level, however, it may be argued that such a strict reliance upon the measured phenomenon as inherently content valid simply regresses to the analysis and arguments already set forth by Banks (1976) and currently under criticism. Rather, within the framework of criterion validity, the phenomenon of white preference as measured within these paradigms may derive its importance from the fact that it concurs with racial self-concepts (as a concurrent criterion) or that it predicts natural events of racial selection

and preference among black children in a real-world domain outside the laboratory (as a predictive criterion).

Empirically, it can be said to do neither. For example, in the study by McAdoo (1970) to which Williams and Morland have referred, black children from the North and the South were measured concurrently for self-concepts and for racial preferences. In this instance, the Williams et al. (1975) PRAM measure of racial preference, in fact, was found to be unrelated to northern black children's self-concepts. And contrary to the assumptions of concurrent-criterion validity that might derive from that instrument, measured white preference among southern black youngsters was directly related not to negative self-concepts but to positive self-concepts.

This is not an isolated example of the lack of criterion validity of preference measures for blacks. Porter (1971) investigated the racial preferences of black and white children and the patterns of their play and friendship preferences in the natural social setting of school. Although children from both samples were found to indicate racially influenced biases in their choice behavior within the laboratory procedure (similar to that used by Williams et al., 1975), Porter reported that "race seemed to play little part in determining the [predictive criterion of actual] friendship patterns" (p. 167).

Construct Validity

It might be argued in response to the points discussed above that the phenomenon of white preference in blacks lacks sufficient conceptual specificity to permit a test of its importance or meaning by so precise a set of content- or criterion-validation hypotheses. We would concur. Were it assumed rather that such measured phenomena represent some underlying hypothetical construct of personality, it would remain to be demonstrated that such an inferred construct has validity. The validation of white preference in blacks as representing the construct, for instance, of negative self-concept could proceed via systematic observations of positive correlations between it and logically convergent behavioral phenomena and negative correlations between its occurrence and that of logically divergent phenomena among blacks. In this regard, one might expect that blacks would with reasonable consistency express a conception of their own qualities and abilities for successful functioning as lower than those of white persons; one might expect, as well, to find that success in such domains as school would relate profoundly to such racially influenced self-assessments among blacks. Certain empirical evidence stands in opposition to both the former (see, e.g., Wylie, 1963; Wylie & Hutchins, 1967) and the latter (Coleman, Campbell, Hobson, McPartland, Mood, Weinfield, & York, 1966; Guggenheim, 1969; Hunt & Hardt, 1969; Wolkon, 1971) notion. Similarly, within a population whose qualities of racial identification and valuation are supposed to work so destructively against a sense of personal optimism and worth, one would hardly expect aspirations toward intellectual/academic excellence and occupational/socioeconomic ascendance to obtain. Yet they do and most often in measure equal to or beyond that of persons whose sense of racial identity ought to place them in a relatively superior position. (See Boyd, 1952; Brook, 1974; Ducette & Wolk, 1973; Gist & Bennett, 1963; Phillips, 1972.)

In summary, scant evidence exists of the tendency of blacks to express preferential evaluative orientations toward white characteristics. Furthermore, the validity of such a phenomenon as a measurement of content or predictive significance for white preference within the real world of social choices, self-esteem, or racial pride seems equally unsupported by empirical evidence; even (or, perhaps, especially) in the case of those examples of research cited by Williams and Morland. Why should we retain for the analysis of behavior in blacks either (a) a conceptual/methodological paradigm in which the null hypothesis pertaining to an empirical phenomenon fails consistently to be rejected or (b) a sense of the importance of a phenomenon whose meaning fails to be verified within any of the existing validation paradigms? Neither the evidence reviewed earlier by Banks (1976) nor that specifically cited by Williams and Morland suggests that we should.

References

Banks, W. C. White preference in blacks: A paradigm in search of a phenomenon. *Psychological Bulletin*, 1976, *83*, 1179–1186.

Boyd, G. The levels of aspiration of white and Negro children in a nonsegregated elementary school. *Journal of Social Psychology*, 1952, *36*, 191–196.

Brook, J. Aspiration levels of and for children: Age, sex, race and socioeconomic status correlates. *Journal of Genetic Psychology*, 1974, *124*, 3–16.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfield, F. D., & York, R. L. *Equality of educational opportunity* (U.S. Department of Health, Education, and Welfare, OE-38001). Washington, D.C.: Government Printing Office, 1966.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281–303.

Ducette, J., & Wolk, S. Locus of control and levels of aspiration in black and white children. *Review of Educational Research*, 1973, *42*, 493–505.

Gist, N., & Bennett, W. Aspirations of Negro and white students. *Social Forces*, 1963, *42*, 40–48.

Guggenheim, F. Self-esteem achievement expectations for white and Negro children. *Journal of Projective Techniques and Personality Assessment*, 1969, *33*, 63–69.

Hunt, D., & Hardt, R. The effect of upward bound programs of attitudes, motivation and academic achievement of Negro students. *Journal of Social Issues*, 1969, *25*, 117–129.

McAdoo, H. P. *Racial attitudes and self-concepts of black preschool children*. Unpublished doctoral dissertation, University of Michigan, 1970.

Phillips, B. School-related aspirations of children with different sociocultural backgrounds. *Journal of Negro Education*, 1972, *42*, 48–52.

Porter, J. D. *Black child, white child: The development of racial attitudes*. Cambridge: Harvard University Press, 1971.

Spencer, M. B., & Horowitz, F. D. Effects of systematic social and token reinforcement on the modification of racial and color concept attitudes in black and white preschool children. *Developmental Psychology*, 1973, *9*, 246–254.

Webster's New Collegiate Dictionary. Springfield, Mass.: G. & C. Merriam, 1974.

Williams, J. E., Best, D. L., Boswell, D. A., Mattson, L. A., & Graves, D. J. Preschool Racial Attitude Measure II. *Educational and Psychological Measurement*, 1975, *35*, 3–18.

Williams, J. E., & Morland, J. K. Comment on Banks's "White preference in blacks: A paradigm in search of a phenomenon." *Psychological Bulletin*, 1979, *86*, 28–32.

Wolkon, G. African identity of the Negro American and achievement. *Journal of Social Issues*, 1971, *27*, 199–211.

Wylie, R. Children's estimates of their schoolwork ability as a function of sex, race and self-esteem level. *Journal of Personality*, 1963, *31*, 204–224.

Wylie, R., & Hutchins, E. Schoolwork-ability estimates and aspirations as a function of socioeconomic level, race and sex. *Psychological Reports*, 1967, *21*, 781–808.

Received June 15, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 37–46

Using Quasi *F* to Prevent Alpha Inflation Due to Stimulus Variation

John L. Santa, John J. Miller, and Marilyn L. Shaw
Rutgers—The State University

The nominal alpha level may be very inflated in much of the published literature where the conventional *F* test is used. This alpha inflation is often caused by ignoring stimulus variation or treating it as a fixed effect. The present article illustrates this problem in a variety of areas and discusses the use of Quasi *F* ratios as a means of achieving generality over both subjects and stimuli. Monte Carlo experiments are reported that examine the performance of the Quasi *F* in a variety of realistic situations in which the data violate distribution and homogeneity of variance assumptions. In general, the Quasi *F* has proved to be robust.

It has long been traditional in psychology to treat subjects as a random effect in analysis of variance (ANOVA). The rationale is simple: For each experiment, we select from the population a small sample of subjects, but we want our results to generalize beyond this sample. Rarely are we interested in presenting our data as being pertinent only to the particular individuals studied. Both Clark (1973) and Coleman (1964) before him have argued that the logic that compels us to treat subjects as a random effect might also lead us to treat stimulus variation as a random effect; that is, in many situations psychologists should seek generality beyond both the stimuli and subjects of an experiment. This reasoning led Clark to recommend an ANOVA model in which both subjects and stimulus items are treated as random effects. Such models are perfectly possible to create, but have the disadvantage that many hypotheses can no longer be tested using the standard *F* distribution. Instead, models with two random effects often necessitate the use of Quasi *F* ratios, which permit one to form analytic tests on the assumption of the standard linear model. Unfortunately, the exact distribution of the Quasi *F* statistic is unknown. However, when degrees of freedom are appropriately adjusted, the conventional *F* distribution can be used to approximate the Quasi *F* statistic (see Winer, 1971, p. 377).

Clark applied the Quasi *F* analysis to a large body of semantic memory research and demonstrated that treatment effects can be significant when stimulus variation is not considered but nonsignificant when both stimuli and subjects are treated as random effects. Following the publication of Clark's article, the Quasi *F* analysis has become common, in fact, almost obligatory for research in semantic memory.

Clark's arguments about stimulus variation and the Quasi *F* have, however, been virtually ignored in other areas of psychology. In the present article, we argue that the Quasi *F* analysis is appropriate to many areas of psychological research and that its use should not be restricted to psycholinguistic investigation of semantic memory. Furthermore, we demonstrate that the Quasi *F* is quite robust with respect to violations of distribution and heteroscedasticity assumptions.

Applicability of the Quasi *F*

The Quasi *F* analysis is potentially applicable in any experiment that employs a sam-

---

Requests for reprints should be sent to John L. Santa, Department of Psychology, Douglass College, Rutgers—The State University, New Brunswick, New Jersey 08903.

ple of stimuli, items, or materials drawn from a larger possible population. Exceptions to this exist whenever stimulus variation is totally confounded with subject variation or in case study experiments of a particular item or stimulus set. Excluding these exceptions, psychologists are usually at least implicitly concerned with generality beyond the sample of materials they happen to be using. Most experiments would be considered uninteresting if their results were obtainable for only one subject or even one sample of subjects. Similarly, results that obtain with only one sample of stimulus materials are of questionable interest in psychology. Consequently, psychologists should be as concerned with generality over stimulus materials as they are with generality over subjects.

A few examples might serve to emphasize the variety of situations in which a Quasi *F* analysis is potentially appropriate. Consider first the situation Clark used to illustrate the importance of item variation. Suppose an experimenter wanted to test a theory claiming nouns are more perceptible than verbs. He or she might proceed by selecting a sample of subjects and assessing each subject's perceptibility threshold for a set of 10 nouns and a set of 10 verbs. The experimenter would then calculate the average threshold for nouns and verbs for each subject and would find that every subject exhibited a lower threshold for nouns than for verbs. The original hypothesis that nouns are more perceptible than verbs would appear to be supported by the data, but few psychologists would be convinced by this experiment. After all, the result might only hold for the single sample of 10 nouns and verbs. In fact, the same result could be obtained if only 1 noun were more perceptible than 1 particular verb. So one sees in this simple case that item variation is obviously important whenever items are confounded with (nested within) treatments. Such situations appear in many areas of psychology other than language research.

For example, a number of psychologists have recently been interested in whether human faces are perceived differently from other objects such as buildings. Such experiments compare reaction time or response thresholds to an arbitrarily chosen set of building pictures versus a set of face pictures. Clearly, individual stimulus pictures are nested within treatment condition, and treatment differences could arise simply from differences in particular item perceptibility. A similar situation exists for a clinical psychologist interested in the effects of alcohol on homosexual and heterosexual arousal. A typical experiment might involve a group of subjects given alcohol and a control group given soda water. Subjects might then view a set of slides, each slide depicting a homosexual or heterosexual act. The dependent variable in such an experiment might be the amount of time spent viewing each slide. Again, conclusions regarding the effects of alcohol on type of sexual arousal would be confounded by individual stimuli nested within the homosexual and heterosexual slide groups–confounded in the sense that the main effect of slide type or the interaction with alcohol could arise from one slide or a small number of slides.

Consider next a social psychologist interested in the effects of sex bias in advertising. A typical experiment might involve groups of males and females reading and evaluating a set of job advertisements. The set of advertisements might contain three types: biased for females, biased for males, and neutral. Assuming there were several ads within each of these classes, one again has a situation in which the experimental variables of interest would be confounded with materials. Thus, item variation must be included in the analysis to make the other variables interpretable.

At this point it should be clear that the potential for confounding variables with item variation is present in many areas of psychology and is not at all restricted to psycholinguistic research. In addition to the specific examples we have outlined, it might be noted that at least two other large areas of research should be particularly concerned with the potential confound between experimental variables and stimuli: developmental psychology and studies that employ questionnaires or tests with subscales. Any time a developmental psychologist changes specific materials to suit the age and ability of sub-

jects, he or she entertains the possibility of confounding experimental effects with specific item differences. Similarly, the same confounding is present whenever an educational psychologist or clinician compares people's performance on a pencil-and-paper test. If the scores that enter into the analysis are summed or averaged over items within subscales, then there is a potential confounding of treatments or groups with individual item variation.

All of our examples so far have been concerned with situations in which item variation is directly confounded with treatment variation. Ignoring item variation in these situations leads to results that might not generalize beyond the specific materials of the experiment; and of more importance, the treatment effects might not even generalize across the various stimuli within the experiment. Unfortunately, the problem of item variation confounding treatment effects does not disappear in designs in which stimulus materials cross all treatment conditions. Such experiments are even more common in psychology. For example, a social psychologist might be concerned with the effects of crowding on perceived friendliness. One group of subjects might be jammed tightly into a small room and asked to rate the friendliness of a sample of faces drawn from high school yearbook pictures. A comparison group of subjects might observe the same pictures while seated in a much more spacious, less crowded room. In such a situation the effect of crowding would not be directly confounded with item variation, since both conditions would receive the same items. Effects of crowding would, however, be confounded with the potential interaction of Treatment × Item such that an effect of crowding might be observed on only a small subset of items. If the effect of crowding were obtained for only a few stimuli, it would probably not be of much general interest. In fact, it might well be argued that an effect such as crowding on perceived friendliness would only be of interest if the result were generalizable beyond the idiosyncratic sample of stimuli used in the particular experiment.

Assuming that our examples are sufficient to illustrate the many situations in which stimulus variation can compromise the interpretation of experimental variables, what can be done about this problem? The first step in solving the problem is to include individual stimuli in one's experimental design as opposed to the current practice of summing or averaging over individual stimuli within treatment conditions. The next step is to decide whether to treat the item variations as a fixed or random effect. Obviously, such a decision is not completely arbitrary, but it is beyond the scope of the current article to debate this issue. For our own part, we tend to agree with Clark (1973, 1976) that stimulus variation should be considered a random effect whenever stimuli are arbitrarily drawn from a potentially larger population and whenever it is desirable to generalize the results beyond the set of stimuli used. In other words, *random* for stimuli should mean about the same thing as *random* does for selecting subjects. (For other sides of this issue consult recent articles by Cohen, 1976; Keppel, 1976; Smith, 1976; and Wike & Church, 1976.)

It is important to emphasize that the experimenter's decision to treat stimulus variation as fixed or random has very important consequences for interpretation of the experiment. Consider, for example, an experiment with I treatments, K subjects, and J items nested within treatments. The design for such an experiment is summarized in Table 1 together with the expected values of the mean squares assuming that items are either a fixed or a random effect. The important point to note is that if item variation is a random effect, then the observed estimates of treatment variation will include a component due to item variation, namely, $K\sigma^2_{b(a)}$. Thus, assuming item variation is fixed when it is in fact random will almost always lead to an inflated value for F. Moreover, the inflation will be more severe as the number of subjects is increased. Even replicating the experiment with new subjects and items will not help, since each experiment will produce an inflated F value. In fact, the inflation can be so large that the actual probability of obtaining the observed F can be 40 to 50 times larger than the nominal alpha level, even with only 10 subjects (see Forster & Dickin-

Table 1
Analysis of Variance Table for Experiment Used in Monte Carlo Runs

Source	*df*	Expected *MS*	
		Items fixed	Items random
Treatments (A)	$I - 1$	$J\sigma^2_{ac} + JK(\sigma^2_a)$	$\sigma^2_{b(a)c} + J\sigma^2_{ac} + K\sigma^2_{b(a)} + JK(\sigma^2_a)$
Items within treatments (B(A))	$I(J - 1)$	$\sigma^2_{b(a)c} + K[\sigma^2_{b(a)}]$	$\sigma^2_{b(a)c} + K\sigma^2_{b(a)}$
Subjects (C)	$K - 1$	$IJ\sigma^2_c$	$\sigma^2_{b(a)c} + IJ\sigma^2_c$
A × C	$(I \times 1)(K - 1)$	$J\sigma^2_{ac}$	$\sigma^2_{b(a)c} + J\sigma^2_{ac}$
B(A) × C	$I(J - 1)(K - 1)$	$\sigma^2_{b(a)c}$	$\sigma^2_{b(a)c}$
Total	$IJK - 1$		

son, 1976). In other words, by randomly sampling stimulus materials and then ignoring stimulus variation or treating it as fixed, it is very easy to obtain, inappropriately, significant treatment effects. Conversely, if item variation is not random with respect to treatment conditions, then the Quasi *F* will tend to provide an inappropriately conservative estimate of the true treatment effect. Such an estimate will become more conservative as the number of items within treatments is increased. Thus, the decision to treat item variation as fixed or random is very important and must be consistent with the true state of affairs in the experiment if the resulting statistic is to be meaningful. It is not simply a question of the experimenter deciding to accept more or less generality. Rather, it is a question of choosing a statistical model appropriate to the experiment.

Robustness of the Quasi F

A practical question remains with respect to the use of the Quasi *F* ratio. Does the *F* distribution provide an adequate approximation of the Quasi *F*—one that is useful in realistic experimental situations? There are already several simulations (Davenport & Webster, 1973; Forster & Dickinson, 1976) that suggest the *F* is an acceptable approximation of the Quasi *F* as long as the total degrees of freedom are sufficiently large (> 18). The Satterwaithe (1946) approximation of degrees of freedom for the Quasi *F* leads to a markedly conservative test only if the stimulus and Stimulus × Subjects variance components are quite small. Thus, the Quasi *F* ratio appears to be a reasonable statistic when the normal model holds. However, what happens when one applies the Quasi *F* in the messy world of real data? The ANOVA model has been useful in psychology largely because it is generally robust under violations of the normative assumptions. Psychologists commonly analyze data in which the cells of the design do not have equal variance or the dependent variables are not normally distributed, for example, latency data, percentage data, or situations in which subjects are either correct or incorrect (0–1).[1]

The present article investigates the performance of the Quasi *F* for these typical violations of the ANOVA model. Since the numerator and denominator of the Quasi *F* ratio are linear combinations of estimated mean squares for several sources in the analysis, one might suspect that the violations typically found in psychological data may have serious effects on the accuracy of nominal alpha levels with the Quasi *F*. In particular, we were interested in the possibility that the Quasi *F* ratio would be an extremely conservative test statistic when errors are not normally distributed and when item variances are heterogeneous.

Consequently, we undertook several Monte Carlo studies of the Quasi *F* under a variety of violations of the normal-model assump-

[1] It should be noted that the 0–1 situation is usually avoided by collapsing over a number of stimulus items to obtain a measure with less restricted variation. Use of the Quasi *F* makes it necessary to enter each subject's score on each item, leading to binary scores for individual entries.

tions. The experimental design we used is given in Table 1. In this design 10 subjects were administered each of 12 items from each of four treatment groups.

The data used in the Monte Carlo experiments varied in three ways: in the distribution of data, in the heterogeneity of item variance, and in the combination of values for the variance parameters $\sigma^2_{b(a)}$, σ^2_c, σ^2_{ac}, and $\sigma^2_{b(a)c}$. The experimental data were drawn from distributions that were normal, exponential, log uniform, binary, or log normal. The combinations of item variance within treatment groups were either homogeneous (1,1,1,1) or heterogeneous in the following ratios for the four treatment groups: (1,2,2,5), (1,2,5,10), and (1,2, 5,20). Finally, the basic parameters of the ANOVA model, $\sigma^2_{b(a)}$, σ^2_c, σ^2_{ac}, and $\sigma^2_{b(a)c}$, were varied to approximate a variety of realistic experimental situations (see Table 2). These parameter combinations can be divided into two basic groups: those for which the item and subject variances are large relative to Subject × Item and Treatment × Subject variances (Cases 1–4 in Table 2) and those for which subject variance and Item × Subject variance are relatively large (Cases 6–8 in Table 2). For Case 5, these variances are all equal.

Table 2
Combinations of Variance Parameters Used in Monte Carlo Studies

Case	$\sigma^2_{b(a)}$	σ^2_c	σ^2_{ac}	$\sigma^2_{b(a)c}$
1	9	9	4	2
2	9	9	2	4
3	9	9	4	4
4	9	9	2	2
5	9	9	9	9
6	3	18	1	38
7	2	25	1	28
8	2	12	1	22

Every trial in the Monte Carlo was conducted as follows: (a) values for the random variables, item, subject, Subject × Item, and error were generated using pseudorandom number generators; (b) these components were then added together with the treatment effect and the overall mean to form an observed data point; (c) this set of data was then submitted to an ANOVA; and (d) the probability of obtaining the observed Quasi *F* value or greater for testing treatment effects was recorded using the appropriate degrees of

Table 3
Observed Proportions of Rejections (1,000 Runs) of Quasi F *in Null Case With Normal Errors*

Case	Variance ratio (1,1,1,1)	(1,2,2,5)	(1,2,5,10)	(1,2,5,20)	*M*
			$\alpha = .05$		
1	.043	.047	.048	.068	.052
2	.051	.052	.051	.073	.058
3	.051	.054	.066	.060	.058
4	.045	.049	.057	.075	.057
5	.045	.046	.051	.048	.048
6	.028	.027	.033	.028	.029
7	.035	.028	.036	.035	.034
8	.032	.042	.031	.036	.035
			$\alpha = .01$		
1	.014	.013	.012	.019	.015
2	.011	.013	.009	.024	.014
3	.012	.017	.014	.019	.016
4	.009	.013	.016	.026	.016
5	.011	.012	.011	.014	.012
6	.004	.005	.003	.003	.004
7	.008	.006	.005	.003	.006
8	.005	.008	.004	.006	.006

freedom. At the end of 1,000 experiments, the proportions of rejections were tabulated for alpha levels of .05 and .01. Under the null hypothesis, the observed proportions of rejections should be very close to these nominal alpha levels. In our discussion below we follow Clark's (1973) notation so that Quasi $F = (MS_A + MS_{B(A)C})/(MS_{AC} + MS_{B(A)})$, $F_1 = MS_A/MS_{AC}$, and $F_2 = MS_A/MS_{B(A)}$.

Table 3 shows the performance of the Quasi F under various violations of the homogeneity of variance assumption. The first column of Table 3 illustrates the rejection rate of the Quasi F in the situation of homogeneous item variances. In general, the Quasi F performs quite well under the normal model but tends to be slightly conservative. Columns 3–5 of Table 3 show the performance of the Quasi F under various violations of the homogeneity of variance assumption up to a 20:1 inequality among treatment variances. For Cases 1–4 in which item and subject variances are large relative to the interaction components, there is a trend for heteroscedasticity to give slightly more rejections than the nominal alpha, but for Cases 6–8 heteroscedasticity makes the statistic slightly conservative. Overall, the Quasi F appears quite acceptable in the face of rather severe departures from the homogeneity of variance assumptions.

Next, we examined the Quasi F in situations in which the data were drawn from a variety of common distributions. Monte Carlo results of the Quasi F for the eight combinations of variance parameters and five types of distributions are presented in Table 4. Inspection of the table reveals that the Quasi F is conservative although not painfully so. When item and subject variances are relatively large (Cases 1–4), the Quasi F tends to be very slightly conservative; it is somewhat more conservative in Cases 6–8 in which subject and Item × Subject variances are relatively large.

Averaging over all cases and looking at the mean rejection rates reveals that nonnormality makes the Quasi F somewhat more conservative than it is with the normal distribution. The best performance is obtained with the normal and log-uniform distributions, the worst with the log-normal distribution. The exponential and 0–1 distributions tend to be intermediate. Among these distributions

Table 4
Observed Proportions of Rejections (1,000 Runs) in Null Case for Quasi F *for Five Error Distributions*

Case	Normal	Exponential	Log uniform	Log normal	0–1
		$\alpha = .05$			
1	.043	.042	.050	.034	.047
2	.051	.039	.048	.021	.040
3	.051	.046	.057	.035	.039
4	.045	.038	.053	.034	.035
5	.045	.037	.038	.030	.046
6	.028	.036	.031	.021	.021
7	.035	.036	.040	.019	.028
8	.032	.028	.046	.023	.023
M	.041	.038	.045	.027	.035
		$\alpha = .01$			
1	.014	.007	.008	.003	.009
2	.011	.006	.008	.007	.005
3	.012	.007	.014	.004	.003
4	.009	.006	.012	.004	.005
5	.011	.011	.012	.007	.008
6	.004	.009	.003	.002	.002
7	.008	.004	.005	.001	.002
8	.005	.003	.005	.004	.005
M	.093	.007	.008	.004	.005

the log normal is the closest in approximating the shape of reaction time distributions typically observed in psychological data; consequently, it may be advisable to take logs of reaction times to normalize the data. Finally, it should be noted that although the Quasi *F* performs well, on occasion the observed rejection rate is only one tenth of the nominal alpha. The relative discrepancy between real and nominal alpha is of course greater when the nominal alpha is smaller.

The results of our simulations make it clear that the Quasi *F* is a useful and robust test. Even at its worst, the test is not outlandishly conservative. For example, a test using a nominal alpha of .05 would almost certainly have a real alpha between .02 and .06. For most researchers this level of uncertainty about a decision is perfectly acceptable.

Power

Let us briefly consider the question of power for the Quasi *F* statistic as contrasted to the *F* statistic obtained by assuming stimulus variation to be a fixed effect. It is, of course, difficult to compare directly the power of these tests unless one assumes a constant experimental circumstance. In other words, assume a situation in which there is a real difference underlying the treatment conditions and an experiment with *K* subjects and *J* items. The question of interest is how do the Quasi *F* and the *F* differ in their ability to detect the difference.

It is relatively easy to evaluate power for a conventional *F* by calculating a noncentrality parameter (NCP) from

$$\mathrm{NCP}(F_1) = 2Kd^2/\sigma^2_{\mathrm{ac}},$$

where d = difference in treatment means. Given the NCP and the degrees of freedom for the experiment (in our situation, 3 and 27), it is possible to look up the theoretical probability of detecting the observed treatment effect using tables of the noncentral *F* (Tiku, 1967). The situation is not so straightforward with the Quasi *F*. However, it is possible to obtain a noncentrality parameter by using

$$\mathrm{NCP}(\text{Quasi } F) = 2JKd^2/[K\sigma^2_{\mathrm{b(a)}} \times J\sigma^2_{\mathrm{ac}} + \sigma^2_{\mathrm{b(a)c}}].$$

The degrees of freedom can then be estimated by the observed average for the 1,000 simulated experiments. Using these estimates

Table 5
Observed Versus Theoretical Power of the Quasi F

	Quasi *F*			F_1	
d	Noncentrality parameter	Theoretical p	Observed p	Noncentrality parameter	Theoretical p
		Case 3: $\sigma^2_{\mathrm{b(a)}} = 9$, $\sigma^2_{\mathrm{ac}} = 4$, $\sigma^2_{\mathrm{b(a)c}} = 4$			
.00	.0	.05	.051	.0	.05
.77	1.0	.10	.109	3.0	.27
1.54	4.0	.34	.324	11.8	.76
2.46	10.2	.75	.734	30.3	.99
3.08	16.0	.92	.914	47.3	.99+
4.00	27.0	.99	.992	80.0	.99+
		Case 7: $\sigma^2_{\mathrm{b(a)}} = 2$, $\sigma^2_{\mathrm{c}} = 25$, $\sigma^2_{\mathrm{ac}} = 1$, $\sigma^2_{\mathrm{b(a)c}} = 28$			
.00	.0	.05	.035	.0	.05
.75	2.3	.14	.145	11.3	.27
1.23	6.0	.40	.408	30.0	.76
1.79	12.8	.80	.790	64.0	.99
2.24	20.0	.97	.937	100.0	.99+
2.91	33.8	.99	.997	169.0	.99+

Note. Difference in treatment means = d; means in the four groups were $4 - d$, 4, 4, $4 + d$; $\alpha = .05$.

one can again consult Tiku's tables to obtain an estimate of the power of the Quasi F. Table 5 illustrates the observed and theoretical power of the Quasi F as contrasted to the power of F_1 for two representative conditions of the ANOVA parameters (Cases 3 and 7).

Two conclusions are apparent from the table. First, the theoretical power of the Quasi F is in good accord with the observed probability of rejection. Again the Quasi F is a reasonable statistic, this time in terms of power. The second conclusion is somewhat more troublesome although quite expected: The Quasi F has considerably less power to detect a given treatment effect than does F_1. One should remember that the columns of Table 5 illustrate different situations. The powers given for Quasi F are those when item variation is in fact random; the powers for F_1 are those when item variation is in fact fixed. Inspection of the expected mean squares in Table 1 reveals that if items within treatments are random it will be more difficult to discern treatment differences than if items are fixed.

Clark (1973) has suggested the use of the statistic $F'_{\min}$ whenever mean squares of the Quasi F are difficult to compute—for example, when the data contain missing observations. For the design discussed in the present article, $F'_{\min}$ is as follows:

$$F'_{\min}(i, j) = MS_{\mathrm{A}}/[MS_{\mathrm{AC}} + MS_{\mathrm{B(A)}}],$$

where i is the number of treatment levels minus one and

$$j = [MS_{\mathrm{AC}} + MS_{\mathrm{B(A)}}]^2 \Big/ \left[\frac{MS^2_{\mathrm{AC}}}{df_{\mathrm{AC}}} + \frac{MS^2_{\mathrm{B(A)}}}{df_{\mathrm{B(A)}}}\right].$$

The statistic $F'_{\min}$ is also equivalent to

$$F'_{\min}(i, j) = \frac{F_1 \cdot F_2}{F_1 + F_2}.$$

Since $F'_{\min}$ is always less than or equal to Quasi F, its computational advantage is to be weighed against the obvious disadvantage of its conservatism. In the recent Monte Carlo study of Forster and Dickinson (1976), the observed alpha level for $F'_{\min}$ was in fact very close to the nominal level. However this study was done with a small range of "nuisance parameter" values and assumed the normal model. Consequently, the generality of their results is limited. The observed alpha level of $F'_{\min}$ for a nominal alpha of .05 in the conditions of our study is presented in Tables 6 and 7. Inspection of the tables shows that $F'_{\min}$ is fairly well behaved for Cases 1 to 5, but is markedly conservative for Cases 7 and 8. Clark (1976) has pointed out one case for which it is possible to compute an exact alpha level for $F'_{\min}$, namely, when there are two treatment groups. In this case, the square root of $F'_{\min}$ is distributed as a t test between two means with unequal variances. In such a situation it is of course perfectly reasonable to use $F'_{\min}$.

Imhoff (1960) generalized Scheffé's two-way mixed model to a completely crossed

Table 6
Observed Proportions of Rejections (1,000 Runs) for F'_{min} *in Null Case With Normal Errors*

	Variance ratio				
Case	(1,1,1,1)	(1,2,2,5)	(1,2,5,10)	(1,2,5,20)	*M*
1	.043	.047	.047	.067	.051
2	.047	.048	.047	.071	.053
3	.049	.051	.064	.056	.055
4	.043	.049	.057	.073	.056
5	.042	.044	.048	.046	.045
6	.008	.009	.009	.008	.009
7	.017	.013	.013	.008	.013
8	.011	.014	.015	.016	.014

Note. α = .05.

Table 7
Observed Proportions of Rejections (1,000 Runs) in Null Case for F'_{min} *for Five Error Distributions*

Case	Normal	Exponential	Log uniform	Log normal	0–1
1	.043	.041	.049	.034	.041
2	.047	.037	.046	.020	.036
3	.049	.045	.054	.033	.035
4	.043	.035	.053	.029	.031
5	.042	.032	.038	.026	.040
6	.008	.013	.007	.006	.004
7	.017	.013	.013	.009	.005
8	.011	.009	.017	.008	.009

Note. $\alpha = .05$.

three-way mixed model with two random effects and one fixed effect and derived an exact test based on Hotelling's T^2 for the hypothesis of no fixed main effects. Again, the test is perfectly appropriate but of limited applicability.

Discussion

Our simulations have shown that the Quasi *F* ratio is a very reasonable statistic. It appears to be slightly conservative, but is robust with respect to violations of the normal model. In short, the Quasi *F* is quite serviceable in realistic experimental situations. There is no doubt that the Quasi *F* is a useful test whenever inferences beyond the specific stimuli and subjects of an experiment are sought. However, some experimenters will be concerned about being penalized by the conservative nature of the test. First, it should be noted that the Quasi *F* is not markedly conservative with respect to the nominal alpha level. Second, experimenters who use the Quasi *F* should feel free to break the stranglehold of $\alpha = .05$. Both experimenters and editors should be willing to accept a result that is significant at the .06, the .07, or even the .10 level. It is far better to relax slightly the nominal level of rejection than to persist in the misuse of a statistic such as F_1 in which the nominal alpha has virtually no meaning.

What about situations in which a fixed effect design is appropriate, for example, when sampling constraints become so stringent that the experimenter cannot treat the stimulus variation as random? Even in these situations the experimenter should not totally ignore stimulus variation. Rather, stimulus variation should still be entered into the design as a fixed effect (as opposed to collapsing over stimuli to obtain an average score for each subject). With this precaution it is at least possible to evaluate the contribution of stimulus variation. In a design with stimuli nested in treatment conditions the experimenter should test for the significance of stimulus variation. When the same stimuli appear in all treatment conditions, the Treatment × Stimulus interaction should be evaluated. If it turns out that stimulus variation or the Treatment × Stimulus variation is not significant, then general confidence in the treatment effect is greatly increased. On the other hand, if these terms are significant, conclusions about the treatment effect might not even generalize across the sample of stimuli used in the experiment. In either case the experiment does not generalize beyond the particular stimulus sample, but if the appropriate item variation term is small, the experimenter can at least have confidence that the treatment effect is not due to a few deviant items.

Another alternative for experiments in which stimuli are a fixed effect is to replicate the experiment using a new sample of subjects and new stimulus materials. Many experimenters already use this procedure. For example, an experimenter might use two lists of words or alternate forms of a test. In such cases it is often possible to use the Quasi *F* test by treating both subjects and replica-

tions as random effects even if the experimenter does not want to analyze responses to individual stimuli. Demonstrating that a result is reliable across replications clearly enhances the credibility of the finding.

The above remarks pertain to a stimulus effect that must be treated as a fixed source of variation. As we have previously noted, it is much more common to want generality across both subjects and stimuli. In such circumstances the Quasi F is both an appropriate and a robust statistic. However, there is a final note of caution: Our discussion of the Quasi F ratio is itself limited in generality. We have examined the performance of only one type of Quasi F statistic. More work is needed to explore the class of Quasi F tests before psychologists proceed to use every imaginable form of concocted F ratio. In the meantime, the robustness of the Quasi F in our simulations is quite promising and suggests that more general use of such tests would greatly increase the reliability of the published data base.

References

Clark, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 335–359.

Clark, H. H. Reply to Wike and Church. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 257–261.

Cohen, J. Random means random. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 261–262.

Coleman, E. B. Generalizing to a language population. *Psychological Reports*, 1964, *14*, 219–226.

Davenport, J. M., & Webster, J. T. A comparison of some approximate F-tests. *Technometrics*, 1973, *15*, 779–789.

Forster, K. I., & Dickinson, R. G. More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and min F'. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 135–142.

Imhoff, J. P. A mixed model for the complete three-way layout with two random-effect factors. *Annals of Mathematical Statistics*, 1960, *31*, 906–925.

Keppel, G. Words as random variables. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 257–266.

Satterwaithe, F. E. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 1946, *2*, 110–114.

Smith, J. E. K. The assuming-will-make-it-so fallacy. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 257–266.

Tiku, M. L. Tables of the power of the F-test. *Journal of the American Statistical Association*, 1967, *62*, 525–539.

Wike, E. L., & Church, J. D. Comments on Clark's "The language-as-fixed-effect fallacy." *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 249–255.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Received July 5, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 47–53

The Detection of Deception

David T. Lykken
Department of Psychiatry
University of Minnesota Medical School

The polygraph (lie detector) test has an accuracy of 64% to 71% (against a chance expectancy of 50%) when the polygraph charts are scored blindly and are thus uninfluenced by clinical impressions of the subject or of the evidence against him. The lie test is biased against truthful subjects, at least half of whom may be erroneously classified as *deceptive*. These conclusions, based on two recent studies of lie test validity in real-life applications, corroborate an earlier critical analysis of the assumptions on which the lie detector is based. Since, in the field, most subjects tend to "fail" the lie test whether they are truthful or deceptive, the method more often detects lying than it does truthful responding. However, it seems probable that deceptive subjects could be taught to artificially augment their polygraph responses to the so-called control questions and thus to avoid being scored as deceptive. The review by Podlesny and Raskin conveys the impression that the lie test is already highly accurate and that addition of other response variables might enhance its validity even further. This impression is erroneous and dangerously misleading.

The recent survey by Podlesny and Raskin (1977) conveys an impression that existing lie detector techniques are based on reasonable psychophysiological theory and are supported by experimental findings of impressive validity. If current techniques already permit valid detection of deception in from 88% to 96% of subjects tested (Podlesny & Raskin, 1977, p. 787), even without the 9 or 10 new variables that the authors describe as "particularly promising" (p. 797), then the further research they call for might be expected to produce a virtually infallible lie detector. Who could then object to the growing trend toward lie detector screening of employees or the admission of lie test evidence in courts (Lykken, 1974)? If the polygraphist is correct 96% of the time even "on hardened criminals behind bars" (Raskin, quoted in Dunleavy, 1976), then allowing a fallible jury to deviate from the polygraphist's judgment—indeed, using a jury at all—can only diminish the accuracy of verdicts based solely on the lie test. In view of the claims that other exponents of the lie detector have been making over the years, it strikes one as curious that such implications of these claims seem never to be spelled out. If the lie test is 96% accurate, at least as Raskin performs it, and if Podlesny and Raskin are correct in suggesting that the few existing studies, most of which are defective in design, have as yet failed to exploit the potential of a host of "promising" additional test variables, it appears that we are dealing not only with the most valid psychological test ever devised but also with potentially the most important social problem ever to issue from the psychological laboratory.

Requests for reprints should be sent to David T. Lykken, Department of Psychiatry, University of Minnesota Medical School, Box 392, Mayo Memorial Building, 420 Delaware Street, S.E., Minneapolis, Minnesota 55455.

Theory of the Lie Detector Test

In an earlier treatment (Lykken, 1974), I attempted to show that the theory of the lie test is so naive and implausible that one should demand especially strong empirical evidence before accepting claims of extremely high validity. Podlesny and Raskin (1977)

Table 1
List of Questions Used in Raskin's Lie Dectector Test (Note 1)

1. Were you born in Hong Hong? (Yes)
2. Regarding the stabbing of Ken Chiu, do you intend to answer truthfully each question about that? (Yes)
3. Do you understand that I will ask only the questions we have discussed? (Yes)
4. During the first 18 years of your life, did you ever hurt someone? (No)[a]
5. Did you cut anyone with a knife on Dumfries St. on January 23, 1976? (No)[b]
6. Before 1974 did you ever try to seriously hurt someone? (No.)[a]
7. Did you stab Ken Chiu on January 23, 1976? (No)[b]
8. Is your first name William? (Yes)
9. Before age 19, did you ever lie to get out of trouble? (No)[a]
10. Did you actually see Ken Chiu get stabbed? (No)[b]

Note. Defendant's answers are in parentheses. If the autonomic disturbance associated with the relevant questions tends to be greater than that associated with the paired control questions, the subject is diagnosed as ***deceptive***. Since it is assumed that an innocent subject is more concerned by the control than by the relevant questions, larger responses to the former are interpreted as evidence that the answers to the latter are ***truthful***.
[a] Control question.
[b] Relevant question.

dismissed my analysis on the grounds that it "assumed that control questions are designed to be answered truthfully by the subject and that a lack of difference in magnitude of reactions to control and relevant questions constitutes the basis for a truthful result. Both of these assumptions are factually incorrect" (p. 787). The second of these alleged assumptions does not appear in my article (Lykken, 1974), which, on the contrary, correctly states that "the examiner is advised to classify tests that give intermediate [i.e., lack of difference] scores as 'inconclusive'" (p. 730). With respect to the first assumption, I stated accurately that "the control question is chosen with the intention that it will elicit an emotional response from the subject, preferably a response involving the attitude of guilt, for example, 'Can you remember ever stealing anything before you were 18 years old?'"; I added, "It is expected that the subject will answer it truthfully" (pp. 729–730). I should have also explained that many polygraphists claim that they can devise control questions, not concerned with the central issue of the interrogation, that the subject will answer deceptively and that will derive their guilty emotional impact from that attempted deception. Instead of somehow invalidating my analysis, this addendum reveals even more clearly the naiveté and implausibility of the theory of this control question lie test, as is illustrated below.

Let us consider an actual lie detector test administered by Raskin to a criminal defendant accused of homicide by stabbing (*Proceedings at Trial*, Note 1). The questions employed in that test are listed in Table 1. Questions 5, 7, and 10 are the relevant questions pertaining to the incident; Questions 4, 6, and 9 are the control questions. The scoring procedure used by Raskin involves comparing the polygraph responses associated with each of the three adjacent pairs of relevant and control questions and assigning a numerical score to each pair, for example, −3 if the relevant question elicits a much larger response than the control, +3 if the control response is much the larger, and 0 if there is no difference. This is done for each of the three or four polygraph channels employed and for each of the two or three repetitions of the question list that may be used. If the sum of these scores is, say, +6 or higher, the subject is said to have been truthful; if the sum is −6 or lower, he is diagnosed as deceptive. In the 10% or so of cases in which the total score is near 0, the test is considered inconclusive.

It should be pointed out that a polygraph chart is very complex and that considerable subjectivity may influence the polygraphist's evaluation of the autonomic disturbance associated with a particular question. But let us assume that objective and consistent rules for evaluating polygraphic response amplitude were available and that some means were found to insure that they were followed faithfully by polygraphists in practice. Then, referring to Table 1, if this defendant tended to give larger autonomic responses to the relevant questions (5, 7, and 10) than he did to the

controls (4, 6, and 9), he would be classified *deceptive*. If his responses to the controls tended to be larger, he would be classified *truthful*. The question is whether it is reasonable to expect that such a test might have 96% validity.

Podlesny and Raskin (1977) explained, referring to the control questions, that "the subject is very likely to be deceptive to them or very concerned about them" (p. 786). I suspect that this defendant's *no* answers to Questions 4 and 9 were technically untrue because most people have "hurt someone" or have "lie[d] to get out of trouble" prior to the age of 20. But I do not know that these answers were false in this case and neither does Raskin. Indeed, it is quite possible that this defendant thought he was telling the truth in both instances, interpreting *hurt* to mean something serious like the mortal stab wound he was accused of inflicting and *trouble* to mean something serious like being charged with murder. In the case of the second control question (Question 6), I think it most likely that the subject's answer was entirely truthful. (I had never tried to "seriously hurt someone" prior to 1974!) A moment's reflection makes it plain that no polygraphist can reasonably claim to be able routinely to construct control questions that are somehow guaranteed to elicit deceptive answers from the subject. If Podlesny and Raskin wish to claim that this is an essential feature of a properly administered lie test and that any analysis of the theory of the test must assume that the control responses are known lies, then further consideration of the lie test is pointless because no such test could possibly be devised except under extraordinary circumstances (viz., if the examiner happens to have proof of additional crimes committed by the subject but which the subject wishes to deny).

Clearly the purpose of the control question is that "it will elicit an emotional reaction from the subject" (Lykken, 1974, p. 730) or that the subject will be "very concerned" about the questions (Podlesny & Raskin, 1977). The emotional reaction occurs because the subect's answer is deceptive, because he is truthful but is ashamed to make that admission, or merely because the question touches on some painful or embarrassing issue. The issue to be decided, then, is whether a larger autonomic response to the relevant question (e.g., "Did you stab Ken Chiu on January 23, 1976?") than to the control question (e.g., "Before 1974 did you ever try to seriously hurt someone?") should plausibly be taken as strong evidence that the subject's answer to the relevant question is a lie.

To state it more generally, does the control question really function as a control in the usual scientific sense of that term, that is, does the autonomic response to the control question provide a reasonable estimate of what the subject's response to the relevant question ought to be if he is answering truthfully? Alternatively, one might ask whether the control response yields a reasonable estimate of what the relevant response should be if the answer to the relevant question is deceptive. This is the basic unavoidable assumption of the lie detector test, and it seems to me to be patently implausible. By what prescience did Raskin know how concerned the defendant was about Question 6? And yet he was obliged to titrate this concern with exquisite precision in advance of the test proper because his scoring assumes that the response to Question 6 will be greater than the response to Question 7 if the answer to Question 7 is truthful but that the response to Question 7 will be greater than the response to Question 6 if the relevant answer is a lie. One can imagine scenarios, all perfectly plausible, that might have led this defendant to be extremely, moderately, or negligibly concerned about the three control questions listed in Table 1. As a general rule, one would expect most subjects to be more concerned about the relevant questions than about the controls, whether they answer deceptively or truthfully, because it is the relevant questions that refer directly to the source of their immediate jeopardy. Thus, one would expect most subjects to tend to "fail" lie detector tests in real-life situations, and this bias against the truthful subject is just what the data confirm, as is shown below.

On the other hand, no psychologist ought to rule out the possibility that some criminal defendants, even though guilty as charged, might develop habituation to, or psychody-

namic defenses against, specific references to their crime and thus might be less responsive to the relevant than to the control questions so as to "pass" the lie test. A sophisticated criminal might know enough to augment his own reactions to the three control questions by flexing his toes, tensing his diaphragm, or biting his tongue at the appropriate moments. No good studies of the success of such countermeasures in real-life situations are available. Polygraphists claim that they could not easily be deceived in this way, whereas, on the contrary, I claim that I could train guilty suspects to successfully "beat" the control question lie test.

In the above case, Raskin testified that the defendant had in fact responded most strongly to the control questions and was truthful. The jury disagreed, finding the defendant to be guilty (*Proceedings at Trial*, Note 1). Did an innocent suspect in this instance behave in accordance with the assumptions of the lie detector or did a guilty suspect beat the test? One must turn to systematic validity studies to determine how the lie test, however implausible, actually works in practice.

Accuracy of the Lie Detector in the Field

Although Podlesny and Raskin (1977) acknowledged that "there are many problems inherent in laboratory investigations of [psychophysiological detection of deception]" (p. 782), they stopped short of asserting that meaningful estimates of lie test accuracy in any field application must be obtained from appropriate field studies. Yet this clearly is the only reasonable conclusion. Giving lie tests to students who have enacted mock crimes or to prisoners who have competed for money prizes may be useful for other purposes but will not provide adequate predictions of what can be expected in real-life criminal investigation. Raskin's laboratory study using prison inmates (Raskin & Hare, 1978), for example, involved a deceptive context in which genuine and realistic fear of failure played no role whatever (surely the fear that one might fail to win a $20 prize is qualitatively different from a criminal suspect's fear that he may end up in prison). Thus, when one looks for evidence concerning the accuracy of the lie test in its intended application, one must confine one's attention to real-life studies in the field.

Secondly, since interest lies in the contribution of the polygraph to the detection of deception rather than in the clinical judgment of the examiner, one must also exclude all studies in which the lie tests were scored globally, still a common practice among many polygraphists. With this method of scoring, all the examiner knows about the subject, the evidence against him, his demeanor during the examination, and the like is compounded in the mind of the examiner with the actual polygraph results by some unspecified subjective formula to produce the final judgment of deceptive or truthful. In the Bersh (1969) study, for example, the criterion of guilt or innocence was the majority verdict of four experienced prosecutors based on their reading of the completed files of 243 criminal suspects. These judges split 2:2 on 27 cases, leaving 216 that could be usefully compared to the polygraphist's previous diagnosis of deceptive or truthful. The polygraphists agreed with the criterion in nearly 88% of these cases. But because at the time of the examination the polygraphists knew what evidence was then available and were able to interview and observe the suspects at some length, one must suppose that their decisions about the suspects' guilt or innocence would have been substantially more accurate than chance expectancy (i.e., 50% agreement with the criterion) even if they had ignored the polygraph results entirely. In fact, one cannot be certain that the polygraph itself contributed at all to the accuracy rate that Bersh reported. Another part of the U.S. Army study from which Bersh's data came ("Use of Polygraphs," 1975) examined the agreement between the original polygraphist's decision and that of other polygraphists who read the same charts blindly. Agreement was low (kappa coefficients ranged from .15 to .51), indicating that the polygraph data could not have contributed much to the accuracy of the original examiner's judgments. Since field studies using blind scoring of polygraph charts do not show nearly so high an ac-

curacy, Bersh's findings must be set aside as ambiguous and almost certainly are an overestimate of the validity of the polygraph test per se.

Clinical Versus Actuarial Lie Detection

Polygraphists who endorse the use of global judgments stress that it is the examiner, not the polygraph, who functions as the lie detector. Trade journals such as *Polygraph* or *Journal of Polygraph Science* are full of unsubstantiated claims concerning interview behaviors said to be indicative of guilt or of innocence (Lykken, 1978). Is it fair, therefore, to insist on separate assessment of the contribution made by the polygraph charts themselves to the accuracy of the examiner's decisions?

The history of validity studies of projective techniques like the Rorschach Inkblot Test provides a limited but useful analogy. It appears that certain talented individuals, observing a subject responding to the Rorschach cards, are often capable of drawing clinical inferences of remarkable accuracy. But the majority of Rorschach test administrators are not nearly so accurate, and attempts to objectify the cues or reasoning employed by the skillful few have met with limited success. There undoubtedly are certain police detectives and polygraph examiners who are similarly skillful in determining, by subjective evaluation of clinical observations, which suspect is lying and which is not. It is possible that more polygraphists might develop such skills if they were given formal training in psychology. But it is most doubtful that a trial judge would ever admit into evidence the clinical opinion of some self-styled "veracity expert," in the form, "I have observed this defendant, considered his story and the relevant evidence, and in my opinion he is (or is not) telling the truth," even if he were a fully accredited psychologist or psychiatrist. What business concern would employ someone who claims to be able to detect lying intuitively, through observing interview behavior, and would allow him to screen prospective employees for honesty or to determine which current employees are stealing from the company and should be fired?

Clearly, the mystique of the lie detector, the reason why the polygraph test is taken seriously by some courts, by business, and by the general public and why the lie detector industry is flourishing in this country, is wholly dependent on the technological or scientific aura of the polygraph itself. It is conceivable that some polygraph examiners are skillful clinical lie detectors, but this issue is of negligible scientific or social importance. What is important is whether, as Podlesny and Raskin contend, the polygraph test is an objective, teachable method of extraordinary validity.

The Evidence

Fortunately, subsequent to my 1974 review, two field studies appeared that provide estimates of lie test accuracy under real-life conditions, when the polygraph charts are scored blindly by someone other than the examiner who administers the test. Both authors were and are professional polygraphists and certainly were not hoping for unfavorable results. Horvath (1977) had 10 trained polygraphists independently score charts taken from the files of a large police department. There were 28 suspects who had subsequently been cleared by the confession of another person; 28 others had themselves confessed some time after the original testing. The 10 experienced polygraphists agreed with each other about 89% of the time, indicating presumably that they followed similar rules of chart interpretation. But the validity of their scoring was not nearly as impressive as their interjudge agreement; the 560 blind scorings were correct only 64% of the time. My analysis of the lie test suggested that it should discriminate against the truthful subject or at least against those subjects with sense enough to realize that the relevant questions are more important to their fates and more threatening than the (essentially irrelevant) control questions. This expectation was strongly confirmed by the Horvath study in which the known liars were correctly scored as deceptive about 77% of the time (against a chance expectancy of 50%), whereas the known truthful suspects were incorrectly

Table 2
Summary of Available Data on Accuracy of Control Question Lie Test Using Blind Scoring

Item	Horvath (1977)	Barland and Raskin (Note 2)	
	Verified	Reported	Corrected[a]
Number guilty	28	40	40
Number innocent	28	11	40
Percent guilty	50%	78%	50%
Percent deceptive	63%	88%	76%
Percent correct (hit rate)	64%	86%	71%
False negative rate	31%	17%	5%
False positive rate	39%	13%	36%
Guilty called truthful	23%	3%	3%
Innocent called deceptive	49%	55%	55%

Note. Observe the high proportion of innocent suspects misclassified as deceptive and the associated high rate of false positive classification.
[a] To provide meaningful estimates of accuracy and error rates, the data of Barland and Raskin had to be corrected for the high base rate (78%) of criterion-guilty subjects. This was done by assuming Raskin would have made the same proportion of errors (55%) if 40 innocent suspects had been tested as he made on the 11 innocent suspects who were tested, thus yielding a standardized base rate of 50% criterion-guilty subjects.

scored as deceptive half of the time, giving a false positive rate of 39%.

The second recent study (Barland & Raskin, Note 2) employed Bersh's method of using a panel of lawyers or judges to determine, from all evidence excluding the lie test results, which suspects were guilty or innocent. Barland administered the tests, and the charts were then scored independently by Raskin. A majority of the criterion judges agreed on 64 of the 92 cases tested, but 13 (20%) of these 64 tests were classified *inconclusive* by Raskin. On the remaining 51 tests, Raskin's scoring agreed with the criterion on 44 of them, a hit rate of about 86%, which the authors reported as their estimate of field accuracy. However, 39, or 78%, of these same cases were guilty by the criterion, which means that one might have achieved a hit rate of 78% on this sample just by calling everyone deceptive (Raskin in fact scored 88% as deceptive; see Table 2). Clearly, nonarbitrary accuracy estimates can only be obtained either by equalizing the numbers of guilty and innocent suspects and assuming Raskin would have been correct in the same proportion of 40 cases as he was in the actual 11 cases (see right-hand column of Table 2) or by considering the fate of the guilty and innocent subjects separately. Raskin scored 39 of the 47 guilty suspects as deceptive, 1 as truthful, and the remaining 7 as inconclusive. But only 5 of the 17 innocent suspects were correctly scored as truthful; 6 were called deceptive and 6 inconclusive. Had this study been designed like Horvath's (1977), with half the subjects guilty and half innocent, then—excluding inconclusives—one might expect about 71% hits overall and a false positive rate of 36%, very similar to the 39% false positives in the Horvath study. Although Raskin correctly diagnosed all but 1 of the guilty subjects as deceptive (not counting the 7 inconclusives), he did this at the expense of calling 55% of the innocent suspects deceptive also.

These two studies constitute the only evidence available concerning the accuracy of the control question lie test administered under real-life conditions and scored to exclude the influence of clinical judgment (or prejudice) and thus to provide some idea of the accuracy of the polygraph test itself. And the two studies agree quite well, showing an accuracy of from 64% to 71%, against a chance expectancy of 50%, and showing that of those who fail the test, 36% to 39% will be false positive, truthful subjects (assuming that half the subjects tested are innocent). Raskin failed a higher proportion of his subjects than Horvath's polygraphists did: 76% versus 63% if one again assumes that both

studies used equal numbers of guilty and innocent subjects. Therefore, Raskin called a higher proportion of both the guilty and the innocent subjects deceptive. Podlesny and Raskin (1977) cited the Barland and Raskin (Note 2) study, but they did not mention the actual results; the Horvath (1977) study was not even cited. Instead, my arguments (Lykken, 1974) predicting that the lie test should show a high rate of false positives in real-life applications are supposedly refuted by a referral to the results of two mock crime laboratory studies by Raskin and his colleagues (Podlesny & Raskin, 1977, p. 787).

Conclusions

Thus one sees that the control question lie test is not 88%, 90%, or 96% accurate in real-life applications, but rather is in the neighborhood of 64% to 71% accurate when standardized for a chance expectancy of 50%. The actual false positive expectancy is not 8%, 4%, or 2%, but is more on the order of 36%–39%. A skillful examiner who is willing to call as many as three fourths of all subjects deceptive can detect most liars (assuming the subjects are not equally skilled at beating the test), but he will at the same time call most of the truthful subjects deceptive also. An interesting research question not mentioned by Podlesny and Raskin is whether many deceptive subjects could be trained to beat the special form of lie test that they advocate. Any intelligent criminal could easily be taught to identify the three control questions and instructed to augment his autonomic reactions to these questions in a variety of covert ways. In fact, I venture another prediction; let me train guilty suspects in Barland and Raskin's next field study, and I predict their false negative rate may approach what their false positive rate is right now.

Reference Notes

1. *Proceedings at trial, Her Majesty the Queen v. William Wong.* Supreme Court of British Columbia, No. CC760628, Vancouver, Canada, October 1976.
2. Barland, G. H., & Raskin, D. C. *Validity and reliability of polygraph examinations of criminal suspects* (U.S. Department of Justice Report No. 76-1, Contract 75-NI-99-0001). Salt Lake City: University of Utah, Department of Psychology, March 1976.

References

Bersh, P. J. A validation of polygraph examiner judgments. *Journal of Applied Psychology,* 1969, *53,* 399–403.

Dunleavy, S. Patty wasn't guilty. *The Star,* December 14, 1976, pp. 24–25.

Horvath, F. S. The effect of selected variables on interpretation of polygraph records. *Journal of Applied Psychology,* 1977, *62,* 127–136.

Lykken, D. T. Psychology and the lie detector industry. *American Psychologist,* 1974, *29,* 725–739.

Lykken, D. T. Uses and abuses of the polygraph. In H. L. Pick (Ed.), *Psychology: From research to practice.* New York: Plenum Press, 1978.

Podlesny, J. A., & Raskin, D. C. Physiological measures and the detection of deception. *Psychological Bulletin,* 1977, *84,* 782–799.

Raskin, D. C., & Hare, R. D. Psychopathy and detection of deception in a prison population. *Psychophysiology,* 1978, *15,* 126–136.

The use of polygraphs and similar devices by federal agencies: Hearings before the Foreign Operations and Government Information Subcommittee of the Committee on Governmental Operations, 93rd Congress, 2nd Session. Washington, D.C.: U.S. Government Printing Office, 1975.

Received August 26, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 54–59

Truth and Deception: A Reply to Lykken

David C. Raskin
University of Utah

John A. Podlesny
Western Reserve Psychiatric Habilitation Center
Northfield, Ohio

In response to Lykken's critique, scientific evidence is presented that shows that control question tests of deception have an accuracy of approximately 90% in the field situation and are highly effective with both innocent and guilty subjects. Lykken's erroneous representation of the theory of such tests is corrected, and his selective and misleading presentation of the scientific data is rectified. The proper interpretation and application of control question tests in the criminal justice context are described, and the tests are shown to be highly beneficial to innocent defendants, the judicial process, and society in general.

In a critique of our recent article (Podlesny & Raskin, 1977), Lykken (1979) attempted to discredit our theoretical analyses and conclusions by using intuitive and speculative arguments and by selective and misleading descriptions of the existing data and literature. Lykken also made bold claims about his ability to train people to "beat" the control question test, and he presented a misleading description of a polygraph test conducted by Raskin in a criminal case. This article attempts to correct those errors with a careful examination of the theory, the scientific data and literature, and the applications of control question tests for truth and deception.

Theory of Control Question Tests

In spite of 5 years of contact with the literature and with concepts of control question tests (see Raskin, 1978), Lykken still does not understand the simple, basic theory. The theory holds that, following a detailed pretest interview, a guilty subject will show relatively stronger autonomic responses to the relevant questions and that an innocent subject will show relatively stronger responses to the control questions, which deal with acts of the same general nature as those covered by the relevant questions. The control question is a stronger stimulus for the innocent subject because he knows he is truthful to the relevant questions; he has been led to believe that the control questions are also very important in assessing his veracity; the manner of explaining the control questions to him and their wording have elicited a *no* answer; and he is either deceptive in his answers, very concerned about his answers, or unsure of his truthfulness because of the vagueness of the questions and problems in recalling the events. His concern about being diagnosed as deceptive produces autonomic reactions to the controls. There is no attempt to "titrate this concern with exquisite precision in advance of the test proper" (Lykken, 1979, p. 49). Control questions are emphasized to all subjects during the pretest interview and immediately following each chart (Podlesny & Raskin, 1978; Raskin & Hare, 1978).

Lykken was simply wrong when he stated that "the theory of the test must assume that the control responses are known

---

Portions of the research reported in this article were supported by Grant FR-07092 from the National Institutes of Health, by Grant 75-NI-99-0001 from the National Institute of Law Enforcement and Criminal Justice, and by Grant 78-NI-AX-0030 from the Law Enforcement Assistance Administration, U.S. Department of Justice. Points of view or opinions stated in this article do not necessarily represent the official position or policies of the U.S. Department of Justice.

Requests for reprints should be sent to David C. Raskin, Department of Psychology, University of Utah, Salt Lake City, Utah 84112.

lies" (p. 49) and that the purpose of the control question is to provide an estimate of the subject's autonomic response to a relevant question answered truthfully. We have never stated that the control question should "function as a control in the usual scientific sense of that term" (Lykken, 1979, p. 49). This statement describes the function of noncritical items in a guilty knowledge test, and it seems to indicate that Lykken does not understand the basic difference between guilty knowledge tests and control question tests. The actual purpose of the control question is to provide a stimulus that will produce a stronger autonomic reaction than the relevant question when the subject is innocent, thereby providing a positive identification of innocent subjects.

Empirical Issues

Our 1977 review (Podlesny & Raskin, 1977) dealt almost exclusively with laboratory research and made suggestions to maximize its generalizability to field applications, but Lykken arbitrarily dismissed the utility of laboratory experiments in estimating the accuracy of field detection of deception. This position betrays a profound lack of understanding of the scientific method and the value of controlled experimentation and diversity of evidence (Hempel, 1966). As we pointed out (Podlesny & Raskin, 1977), the best strategy "is to employ laboratory research that simulates field-deceptive contexts as closely as possible, along with field validation" (p. 784). We have used prisoners, criminal suspects, very realistic mock crimes (so realistic that some subjects decline to participate in the mock crime), substantial motivation, and potential loss of reward or punishment. In laboratory experiments with subjects recruited from the community by newspaper ads, with prison inmates, and with diagnosed psychopaths, we have consistently obtained accuracy rates above 90% (Podlesny & Raskin, 1978; Raskin & Hare, 1978; Rovner, Raskin, & Kircher, Note 1), and such findings are very useful in the scientific enterprise of estimating accuracy in real-life situations.

Table 1
Percentage of Correct Decisions in Five Studies With Blind Interpretations of Polygraph Charts Obtained From Verified Guilty and Innocent Subjects

Study	Guilty	Innocent
Horvath & Reid (1971)	75[a]	83[a]
	89[b]	94[b]
Hunter & Ash (1973)	88	86
Slowik & Buckley (1975)	85	93
Wicklander & Hunter (1975)	95	93
Raskin (Note 3)	93[c]	69[c]
	100[d]	95[d]
Combined results	90	89

[a] Decisions were made by intern examiners.
[b] Decisions were made by experienced examiners.
[c] Evaluation was nonnumerical.
[d] Evaluation was numerical.

Lykken was correct in emphasizing the need for validation studies with criminal suspects using blind evaluation of polygraph charts. Unfortunately, he provided misleading interpretations of the two studies that he selected, and he failed to mention five published studies that meet his criteria but provide strong evidence against his position. Lykken also failed to mention that in the Horvath (1977) study the original examiner was 100% correct and the cases were all verified by confessions. It has been pointed out (Raskin, 1978) that the unusually low level of accuracy attained by Horvath's blind evaluators was very likely due to their lack of formal training in systematic chart interpretation and their heavy emphasis on overt behavior symptoms rather than on systematic chart interpretation. Therefore, the Horvath study is of little value in assessing the accuracy of decisions based on systematic chart interpretation.

Lykken (1979) was incorrect when he stated that the Horvath (1977) and Barland and Raskin (Note 2) studies "constitute the only evidence available concerning the accuracy of the control question lie test administered under real-life conditions and scored to exclude the influence of clinical judgment (or prejudice) and thus to provide some idea of the accuracy of the polygraph test itself" (p. 52). There are five other published

studies that meet Lykken's criteria of blind interpretation of confirmed polygraph charts from criminal suspects. The findings of these studies are presented in Table 1. The mean accuracy rates of 90% correct on guilty suspects and 89% correct on innocent suspects were based on a total of 1,204 independent decisions obtained by blind interpretation of polygraph charts by 55 different polygraph examiners. The data from Horvath and Reid (1971) show that experienced examiners are more accurate, and the Raskin (Note 3) data clearly demonstrate that the use of a relatively objective, systematic method of quantified chart interpretation yields significantly higher accuracy rates, which approach 100%.

The reliability of the numerical scoring system is extremely high. Using the numerical system with blind chart interpretation, we obtained a mean correlation of .86 for the 15 pairings of six independent evaluators (Barland & Raskin, 1975), a .91 correlation between numerical scores and 99% agreement with the examiner's original decisions on 102 criminal suspects (Barland & Raskin, Note 2), a .97 correlation between numerical scores and 100% agreement with the original examiner's decisions in a laboratory study (Podlesny & Raskin, 1978), and 95% accuracy and 100% agreement with decisions made 2 years before in a laboratory study of criminals and psychopaths (Raskin & Hare, 1978).

The extensive and consistent findings just described demonstrate the very high reliability and validity of blind chart interpretation when it is performed by competent examiners who have been adequately trained in chart interpretation and who do not make decisions based on the questionable procedure of observing behavior symptoms. The latter procedure has been shown to be ineffective in assessing truth and deception (Podlesny & Raskin, 1978; Raskin, Barland, & Podlesny, Note 4), and it is not surprising that examiners trained to rely on behavior symptoms instead of polygraph charts produce results hardly better than chance (Horvath, 1977).

In addition to the Horvath (1977) study, Lykken placed great weight on the high false positive rate in the Barland and Raskin (Note 2) study. In this study, as in field studies of lie detection generally, it was necessary to substitute criteria of guilt or innocence in place of factual knowledge. The two major criteria were decisions of a panel of legal experts based on case information (with all references to the polygraph tests deleted) and judicial outcomes. Those criteria failed to provide assessment of accuracy equivalent to that available in laboratory studies or confirmed criminal cases. Raskin (1978) has stated that the panel criterion is open to serious challenge because the information provided in the Barland and Raskin study was generally inadequate, agreement between the court decisions and the panel was less than perfect, and inherent bias may have existed toward judgments of innocence based on the tradition of the assumption of innocence in the absence of extremely strong evidence to the contrary. Furthermore, the number of criterion-innocent subjects was very small. As a result the 95% confidence interval for the false positives was 11%–59%, whereas the larger sample size for guilty subjects yielded a 95% confidence interval of 0%–16% for false negatives. Therefore we consider these data and those of Horvath (1977) to be of relatively low value in contrast to those presented in Table 1 and the other data described earlier that were ignored by Lykken.

Issues in the Application of Control Question Tests

Lykken (1979) stated that a sophisticated criminal might be able to augment his reactions to the control questions and "pass" the test. The only published study with such procedures used the guilty knowledge test (Lykken, 1960), which is more susceptible to false negatives than the control question test (Podlesny & Raskin, 1977) because the guilty knowledge test employs only skin resistance measures, and a truthful outcome does not require larger responses to the noncritical items, as does the control question test. The subjects were medical students, staff psychiatrists, and psychologists who were

given detailed instructions about the test structure, a strategy and methods to beat the test, and biofeedback training to control their skin resistance responses. Even with minimal consequences for being detected, the accuracy was 100%. Lykken's (1960) failure to train sophisticated subjects with little at stake to be able to beat the simpler guilty knowledge test raises extreme doubt concerning Lykken's statement, "I claim that I could train guilty suspects to successfully 'beat' the control question lie test" (p. 50). Rovner et al. (Note 1) are presently engaged in an extensive laboratory study to assess the effects of detailed information and practice on the accuracy of control question tests.

Although we are opposed for a variety of scientific and ethical reasons to the use of polygraph tests in employment situations, Lykken's opposition to the use of polygraph evidence in court is based on his lack of understanding of the control question technique, his highly selective presentation of the scientific evidence, his misinterpretations of those data that he selected for discussion, and his gross misunderstanding of the criminal justice system. The issues surrounding court use of polygraph evidence involve the level of confidence that can be placed in a truthful or deceptive outcome, the way in which such outcomes are used in the criminal justice process, and the impact of such evidence on juries.

We agree that the data indicate that false positives are more likely than false negatives, even though the rates of both types of errors are low. Even if Lykken were correct concerning the rate of false positives, for practical purposes the confidence in a truthful outcome is higher than that in a deceptive outcome, since a truthful result is more likely to be correct than is a deceptive result. The use of such findings coincides with our judicial and moral standards for acquittal and conviction. Because criminal guilt must be demonstrated beyond a reasonable doubt, considerable evidence is required for conviction, and a deceptive polygraph result is far from sufficient. In the absence of other strong evidence of guilt, no competent or ethical prose-

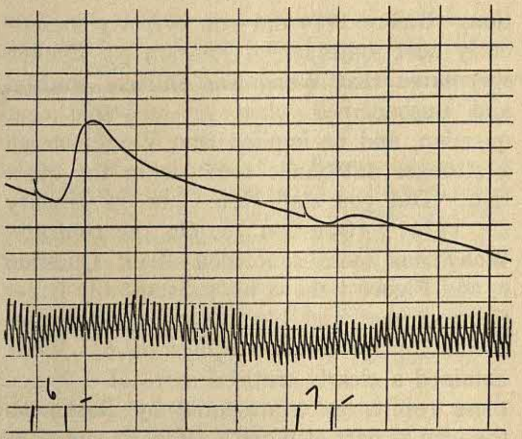

Figure 1. Skin resistance and blood pressure responses of homicide defendant William Wong to a control question (6) and a relevant question (7). The vertical marks indicate the beginning and end of the questions, and the minus sign indicates a *no* answer at that time.

cutor could or would try a case on the basis of a deceptive polygraph test. However, in the absence of overwhelming evidence to the contrary, the high degree of accuracy of a truthful polygraph result should be sufficient to cast the reasonable doubt required for dismissal or acquittal. It has become common practice for law enforcement agencies and prosecutors to dismiss charges in such situations. Given the very low accuracy of some types of evidence, such as eyewitness testimony, that are commonly used against defendants and the great weight accorded to this evidence by prosecutors and juries (Buckhout, 1974), it makes good sense to provide an opportunity for innocent suspects and defendants to clear themselves by means of a properly conducted polygraph test.

Lykken (1979) may have provided a misleading description of the William Wong case (*Proceedings at Trial,* Note 5) by failing to describe its outcome. Wong was accused of homicide on the basis of highly questionable eyewitness accounts. Wong was administered a polygraph test by Sergeant Smith of the Vancouver, Canada police department and was retested prior to his trial by Raskin. He was found to be truthful by both examiners. Using standard techniques, Raskin employed typical control questions, including the ques-

tion, "Before 1974 did you ever try to seriously hurt someone?" (Question 6). Lykken speculated that Wong was entirely truthful and unconcerned when he answered that question, and he implied that Wong showed a stronger autonomic response to the question, "Did you stab Ken Chiu on January 23, 1976?" (Question 7). On the contrary, Wong was more concerned about Question 6, and Figure 1 shows his substantially larger electrodermal and blood pressure responses to the control question (Question 6)! Wong obtained a clearly truthful score of +9, and these results were presented by Raskin to the jury as part of Wong's defense against the murder charge. It should also be mentioned that in the same court hearing, Lykken unsuccessfully opposed presenting to the jury the results of Raskin's and Sergeant Smith's polygraph tests in the defense of innocent homicide defendant William Wong. Lykken's position at the trial was in direct conflict to his previously published position that "judicious use of the polygraph in the criminal investigation context not only can improve the efficiency of police work but could also serve as a bulwark to protect the innocent from false prosecution" (Lykken, 1974, p. 738).

Lykken (1979) also claimed that the use of polygraph evidence in court would overwhelm the jury and might even be used to replace the jury system. His speculations are naive with regard to the judicial process and betray a lack of knowledge of the evidence concerning the impact of the testimony of polygraph experts on jury deliberations. As Tarlow (1975) pointed out, polygraph evidence is simply an aid to the jury in its complicated task of assessing the credibility of witnesses. As such, if the polygraph evidence has probative value, the jury is merely asked to consider it along with the other evidence in the case and to accord it whatever weight the jury finds appropriate. We have not suggested that polygraph tests should replace the jury system. In fact, the available evidence (Tarlow, 1975) indicates that juries are very cautious and that they tend to be "underwhelmed" by polygraph testimony.

Conclusion

The results of many scientific studies in laboratory and field settings as well as our published report to the U.S. Department of Justice (Raskin et al., Note 4) indicate that the accuracy and reliability of control question tests can be very high. On the basis of the present evidence, it is reasonable to conclude that the results of control question polygraph examinations conducted by competent and ethical examiners can have important and beneficial effects for the criminal justice process and for our society in general.

Reference Notes

1. Rovner, L. I., Raskin, D. C., & Kircher, J. C. *Effects of information and practice on detection of deception.* Paper presented at the meeting of the Society for Psychophysiological Research, Madison, Wisconsin, September 1978.
2. Barland, G. H., & Raskin, D. C. *Validity and reliability of polygraph examinations of criminal suspects* (U.S. Department of Justice Report No. 76-1, Contract 75-NI-99-0001). Salt Lake City: University of Utah, Department of Psychology, March 1976.
3. Raskin, D. C. *Reliability of chart interpretation and sources of errors in polygraph examinations* (U.S. Department of Justice Report No. 76-3, Contract 75-NI-99-0001). Salt Lake City: University of Utah, Department of Psychology, June 1976.
4. Raskin, D. C., Barland, G. H., & Podlesny, J. A. *Validity and reliability of detection of deception* (U.S. Department of Justice Final Report, Contract 75-NI-99-0001). Salt Lake City: University of Utah, Department of Psychology, August 1976.
5. *Proceedings at trial, Her Majesty the Queen v. William Wong.* Supreme Court of British Columbia, No. CC760628, Vancouver, Canada, October 1976.

References

Barland, G. H., & Raskin, D. C. An evaluation of field techniques in detection of deception. *Psychophysiology,* 1975, *12,* 321–330.

Buckhout, R. Eyewitness testimony. *Scientific American,* 1974, *231*(6), 23–31.

Hempel, C. G. *Philosophy of natural science.* Englewood Cliffs, N.J.: Prentice-Hall, 1966.

Horvath, F. S. The effect of selected variables on interpretation of polygraph records. *Journal of Applied Psychology,* 1977, *62,* 127–136.

Horvath, F. S., & Reid, J. E. The reliability of polygraph examiner diagnosis of truth and deception.

Journal of Criminal Law, Criminology, and Police Science, 1971, *62*, 276–281.

Hunter, F. L., & Ash, P. The accuracy and consistency of polygraph examiners' diagnoses. *Journal of Police Science & Administration*, 1973, *1*, 370–375.

Lykken, D. T. The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, 1960, *44*, 258–262.

Lykken, D. T. Psychology and the lie detector industry. *American Psychologist*, 1974, *29*, 725–739.

Lykken, D. T. The detection of deception. *Psychological Bulletin*, 1979, *86*, 47–53.

Podlesny, J. A., & Raskin, D. C. Physiological measures and the detection of deception. *Psychological Bulletin*, 1977, *84*, 782–799.

Podlesny, J. A., & Raskin, D. C. Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology*, 1978, *15*, 344–359.

Raskin, D. C. Scientific assessment of the accuracy of detection of deception: A reply to Lykken. *Psychophysiology*, 1978, *15*, 143–147.

Raskin, D. C., & Hare, R. D. Psychopathy and detection of deception in a prison population. *Psychophysiology*, 1978, *15*, 126–136.

Slowik, S. M., & Buckley, J. P. Relative accuracy of polygraph examiner diagnosis of respiration, blood pressure, and GSR recordings. *Journal of Police Science & Administration*, 1975, *3*, 305–309.

Tarlow, B. Admissibility of polygraph evidence in 1975: An aid in determining credibility in a perjury-plagued system. *Hastings Law Journal*, 1975, *26*, 917–974.

Wicklander, D. E., & Hunter, F. L. The influence of auxiliary sources of information in polygraph diagnoses. *Journal of Police Science & Administration*, 1975, *3*, 405–409.

Received August 1, 1978 ■

Notice on Author Alterations

Effective January 1979, the procedure for billing authors for alterations will change. Authors of articles in APA journals are customarily charged for alterations on proofs that result from (a) authors' errors or omissions in the manuscript and (b) changes that result from authors' failure to review the edited manuscript or the proof. Until now, alteration charges have been determined by the number of lines affected by changes requested and have been based on 1974 printing rates. Beginning with the proofs for the first issues of 1979, authors' alterations will be determined not only by the number of lines affected but also by the number of figures and pages affected and will be based on current printers' rates.

This change in alteration billing is being made so that alteration charges will more closely reflect the actual cost of changes requested and may mean larger alteration bills for some authors. Authors are reminded that alteration charges apply only when changes are made on proofs and that *there is no charge* for changes made at the manuscript stage.

Psychological Bulletin
1979, Vol. 86, No. 1, 60–66

Using Distance Information in the Design of Large Multidimensional Scaling Experiments

Jed Graef
University of Toronto, Toronto, Canada

Ian Spence
University of Western Ontario, London, Canada

The principal objective of this study was to discover which distances are most important in determining the recovery performance of a nonmetric multidimensional scaling algorithm. Using Monte Carlo methods we show that the large distances are critical to satisfactory performance, whereas the small and the medium distances play a much less crucial role. This finding has been reliably demonstrated across a variety of conditions, although only for a single combination of dimensionality and number of points. It turns out that certain parallels exist between this work and previous results obtained using cyclic and other incomplete designs. Finally, on the basis of these results we make some recommendations to experimenters regarding data collection procedures; these represent a simple alternative to the methods advocated by Spence and Domoney.

Good nonmetric multidimensional scaling algorithms have been available for more than a decade (Kruskal, 1964), and the large number of published applications attests to their popularity and usefulness. These procedures are capable of constructing a configuration of points in a metric space using no more than the ordinal properties of the data in a matrix of dissimilarities—thus the characterization "nonmetric." Despite the fact that a large body of experience and knowledge has been accumulated regarding the behavior of nonmetric algorithms, there have been few systematic attempts to discover which characteristics of the input data are essential for successful construction of the configuration. It is important to know the answer to this question, since in many situations it may not be feasible, or desirable, to collect all possible pairwise judgments of dissimilarity from a subject. The most obvious situation is one in which the number of stimuli is large and the resulting number of potential paired comparisons becomes too onerous a burden for even the most dedicated subject. Consequently, if some incomplete fraction of the data is to be obtained, it is essential that information be gathered regarding the distances that are most influential in determining the solution. In this article we use Monte Carlo techniques to discover which distances are most important in determining the nature of the final configuration of points and on the basis of our results suggest some possible data collection procedures for the experimenter who wishes to use a large number of stimuli.

This article is based on a paper presented at the annual meeting of the Psychometric Society, Bell Laboratories, Murray Hill, New Jersey, April 1976. This study was supported by National Research Council of Canada Grant A8531 to the second author.

Requests for reprints should be sent to Ian Spence, Department of Psychology, University of Western Ontario, London, Ontario, Canada N6A 5C2.

Method

The design of the study is straightforward and is summarized in Figure 1. A two-dimensional configuration with 31 points was generated randomly within the unit circle. Dissimilarities were simulated by computing error-perturbed interpoint distances according to two different models: multidimensional Thurstone Case V (e.g., Ramsay, 1969)—

$$d_{ij}^e = [\sum_{a=1}^{2} (x_{ia}^e - x_{ja}^e)^2]^{\frac{1}{2}}$$

with

$$x_{ia}^e = x_{ia} + N(0, \sigma_R^2); \quad (1)$$

Wagenaar and Padmos (1971)—

$$d_{ij}^e = [\sum_{a=1}^{2} (x_{ia} - x_{ja})^2]^{\frac{1}{2}} \cdot N(1, \sigma_{WP}^2), \quad (2)$$

that is, the true distance is multiplied by a random normal deviate with a mean of one. Although not invented by Ramsay (1969), for convenience we shall refer to the first model as the Ramsay model: This (noncentral χ^2) model produces an error distance whose variance increases as a function of increasing true distance, although the effect is actually quite small. For the Wagenaar–Padmos model, however, the effect is much more dramatic with the standard deviation of the error distance linearly related to the true distance. This error model is not unlike the lognormal error model that Ramsay (1977) has suggested is the most plausible for many psychological situations—the major difference being the skewness of the lognormal. The distributions are illustrated in Figure 2. The standard deviations in panel a and panel b for the Ramsay model are actually different but not by much. However, the standard deviation for panel a in the Wagenaar and Padmos model is larger than the standard deviation in panel b by exactly the same proportion as the ratio of the two true distances. These error distributions model what we feel to be the reasonable extremes that might be encountered with real data.

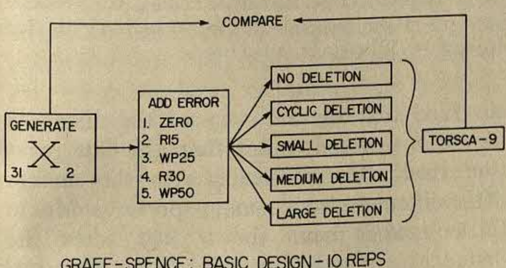

Figure 1. The design of the study. (R15 and R30 refer to the Ramsay model, with $\sigma_R = .15$ and $\sigma_R = .30$, respectively. WP25 and WP50 refer to the Wagenaar and Padmos model, with $\sigma_{WP} = .25$ and $\sigma_{WP} = .50$, respectively. REPS = replications; TORSCA = computer program used.)

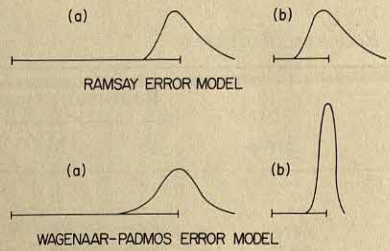

Figure 2. The two error models.

In the sequel, R15 and R30 refer to the Ramsay model with $\sigma_R = .15$ and $\sigma_R = .30$, respectively. These represent medium and high error levels and are the same as those used by Spence and Domoney (1974). For the Wagenaar and Padmos model, WP25 and WP50 refer to situations where $\sigma_{WP} = .25$ and $\sigma_{WP} = .50$, respectively. These two levels were chosen such that the variance of an average distance would be approximately equal in both error models—the appropriate relation was determined algebraically. As will be seen from the results, comparable recoveries were obtained with the two models.

The dissimilarity matrices were analyzed, using TORSCA-9 (Young, Note 1) in five ways—with no deletion, using a maximum efficiency cyclic design[1] with one-third deletion, and after deletion of the small, the medium, and the large distances (a $\frac{2}{3}$ fraction being retained in each case). Thus, for example, when the small distances were deleted, the 155 smallest of the 465 dissimilarities were discarded prior to analysis.

Results

The results for 10 replications are shown in Table 1 and Figure 3. With no deletion, we obtain the same pattern as in many previous Monte Carlo studies and note further that R15 and WP25 were about equivalent in their effects, as are R30 and WP50. This

[1] A cyclic design is a particular kind of partially balanced incomplete block design. It is not central to the substance of this paper that the reader understand how such designs are constructed and used in the multidimensional scaling context. A full and detailed description is given by Spence and Domoney (1974), and it is shown that high efficiency cyclic designs perform very well.

Table 1
Mean Recovery Measures

Error	Root mean square correlations S	M	L	A	Mean absolute error S	M	L	A
				No deletion				
Zero	1000	1000	1000	1000	00	00	00	00
R15	867	786	918	978	07	07	06	07
WP25	901	791	902	980	06	06	06	06
R30	572	522	712	899	13	13	14	13
WP50	659	542	728	919	13	13	14	13
				Cyclic deletion				
Zero	991	982	993	999	02	02	02	02
R15	798	718	872	965	09	09	08	09
WP25	830	689	852	964	08	09	09	09
R30	368	327	496	751	19	22	30	24
WP50	371	326	483	735	19	23	33	25
				Small deletion				
Zero	976	975	996	997	03	02	02	02
R15	792	705	892	964	10	09	08	09
WP25	765	685	825	943	11	10	11	11
R30	312	268	550	704	22	24	36	27
WP50	524	431	639	833	17	17	25	20
				Medium deletion				
Zero	994	986	999	999	02	02	01	01
R15	771	688	857	959	10	10	09	09
WP25	823	678	825	959	09	10	09	10
R30	463	421	555	837	17	19	21	19
WP50	409	375	533	800	19	21	24	21
				Large deletion				
Zero	839	443	−229	505	14	31	78	41
R15	600	355	−091	616	15	27	57	33
WP25	713	355	067	609	13	26	61	33
R30	224	182	095	397	19	41	78	46
WP50	335	112	−102	296	20	48	93	54

Note. S, M, L, and A represent the recovery statistics for small, medium, large, and all distances, respectively. R15 and R30 refer to the Ramsay model, with $\sigma_R = .15$ and $\sigma_R = .30$, respectively. WP25 and WP50 refer to the Wagenaar and Padmos model, with $\sigma_{WP} = .25$ and $\sigma_{WP} = .50$, respectively.

holds for both the root mean square (rms) correlation between true and recovered distances and for the mean absolute error; the latter is simply the average discrepancy between the generated and the recovered distances. Thus, the results seem to be independent of the error process involved. Consequently, since the two error models used are quite dissimilar, it is plausible that comparable results would be obtained using other error distributions. It should also be observed that the recovery of the distances has been assessed rather than the configuration itself. In practical terms this makes little difference, since many previous Monte Carlo studies have shown that when the distances are well recovered, so is the configuration and vice versa. For example, Spence (in press) found a correlation of .967 between recovery of the distances and recovery of the configuration.

In addition to assessing the recovery of

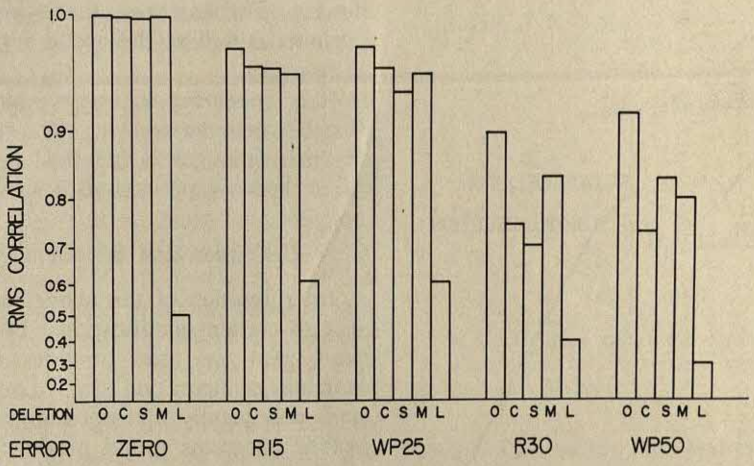

Figure 3. The recovery correlations as a function of zero (0), cyclic (C), small (S), medium (M), and large (L) deletions. (Medium and high error levels were used with both error models, as well as a no-error condition; RMS = root mean square. R15 and R30 refer to the Ramsay model, with $\sigma_R = .15$ and $\sigma_R = .30$, respectively. WP25 and WP50 refer to the Wagenaar and Padmos model, with $\sigma_{WP} = .25$ and $\sigma_{WP} = .50$, respectively.)

the total set of distances, we computed recovery statistics for the small, the medium, and the large distances. These correlations should be interpreted with some care because they are based on subsets of distances with different variances. If allowance is made for these different variances, the correlations for the small, medium, and large subsets are very close in almost all instances. The one striking exception will be discussed presently.

Cyclic deletion produced satisfactory results similar to those obtained by Spence and Domoney (1974) and Spence (in press) and warrant no further discussion. Deleting the small and medium distances produced recovery statistics very similar to the cyclic-design condition, with perhaps slightly better recovery in the high error conditions. However, when the large distances were deleted, the recovery deteriorated dramatically: Even in the zero-error condition the overall recovery was quite clearly unacceptable, and in no condition does the recovery correlation exceed .7. Tschudi (1972) has suggestd that .7 is the minimum acceptable correlation; solutions with recovery correlations smaller than this bear very little resemblance to the generated configurations. The analysis by subsets of distances shows that the large distances are the worst recovered, with some correlations negative and the best around zero. The medium and smaller distances are not as well recovered as in other deletion conditions. Obviously, a nonmetric scaling algorithm performs very poorly when it is denied information relating to the large distances.

In this study we cannot consider the incomplete fractions that were scaled to be experimental designs in the usual sense, since the missing dissimilarities are not determined a priori but are a function of the actual distances. In other words, the situation is not exactly the same as if the pattern of missing distances had been determined independent of the data, as for example, if a cyclic design had been used. Nevertheless, the end result is similar in that an incomplete matrix is scaled. Consequently, it does not seem unreasonable to refer to the pattern of retained dissimilarities as a pseudodesign. The analysis of variance (ANOVA) efficiency cannot be calculated as simply as in the case of cyclic designs (see Spence, in press; Spence & Domoney, 1974), however, it is easy to calculate the mean resistance of the pseudodesigns. This is done by considering the graph of the design to be an electrical network, with the edges of the graph corresponding to conductors of unit resistance. Johnson and Van Dyk (Note 2) have shown

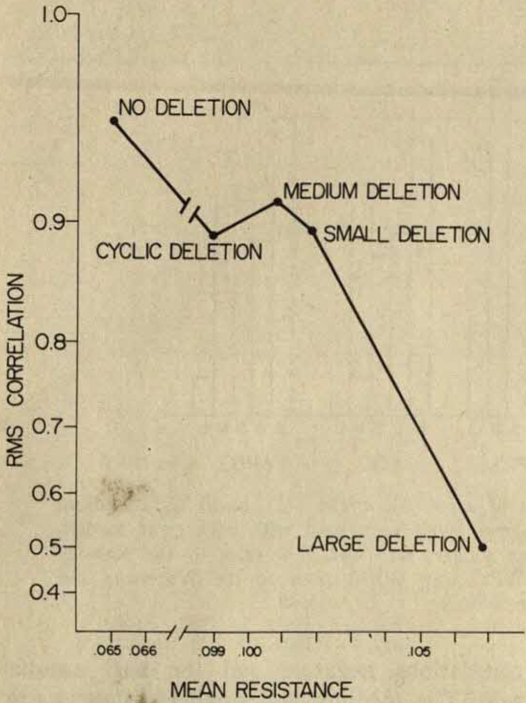

Figure 4. The recovery correlations as a function of the resistance of the pseudodesigns. (RMS = root mean square.)

that mean resistance and ANOVA efficiency are formally identical with resistance inversely proportional to efficiency. We have calculated the mean resistances for all conditions in this study, and the results are shown in Figure 4, where the recovery means are averaged over all error conditions. For the no-deletion and cyclic conditions, the mean resistance is, of course, a constant; whereas for the small-, medium-, and large-deletion conditions the values of mean resistance are the averages over the 10 replications. It can readily be seen that as mean resistance increases, the recovery correlation deteriorates. This is precisely analogous to the results of Spence (in press), where it was found that decreases in the efficiency of cyclic designs were accompanied by deteriorating recovery. It is interesting to note that deleting the large distances produces a pseudodesign with the highest mean resistance. Deletion of small or medium distances produces pseudodesigns with resistances not much higher than that of the maximum efficiency cyclic design. Furthermore, these pseudodesigns perform as well as the cyclic design in terms of recovery.

It is reassuring to observe that the relationship between recovery and resistance for the pseudodesigns is identical to that found for orthodox experimental designs.

Discussion and Recommendations

Interpretation of the above results is subject to certain qualifications. First, the program that we used provides an excellent starting configuration (see Spence, 1972), and consequently it is not known whether all scaling programs would produce comparable results. However, programs that utilize similar methods to compute the initial configuration (e.g., KYST, ALSCAL, SSA-I) should yield good results. Second, in order to conserve computer time, we did not systematically investigate configurations in higher dimensionalities, nor with larger numbers of points. Increases in the values of these parameters would have made the study extremely costly. (We did examine a few individual cases with larger numbers of points and found the same pattern of results.) Thus, extrapolation of our results is not without risk, but we are confident that with larger numbers of points there would be no change in the conclusions (cf. Spence & Domoney, 1974).

In practical terms, if an experimenter is to construct an incomplete design with the small- or medium-dissimilarity comparisons eliminated, some prior information as to the relative sizes of the dissimilarities is necessary. Such information may be obtained very easily; we suggest a number of relatively obvious methods. No doubt the reader can think of others.

1. The experimenter may rank or judge the complete set of dissimilarities. Even if this takes considerable time, the effort will usually not be deemed to be unacceptable, especially in the context of the total time and effort required to plan and set up the average experiment.

2. Pretests with actual subjects may be employed—here there is no compelling necessity for a single subject to perform the whole

task, since only approximate information is required. Several subjects may each judge a subset of the pairs.

3. Pretesting with one or more subjects using the methods of sorting may be used. Subjects will be required to group the stimuli in such a fashion as to maximize within-group similarity and minimize between-group similarity; the number of groups (k) may be fixed in advance or left to the subject but should in either case be small. One will subsequently choose not to collect dissimilarity judgments for the within-group pairs. Unfortunately, if k is much larger than three or four, this method will not yield a sufficiently large number of pairs to be discarded. The situation may be improved, to some extent, by using several judges, since their groupings will not, in general, be identical.

Independent of the strategy employed to decide which dissimilarities are to be collected, it is imperative that a sufficient number of judgments be obtained from the subject. Too few data values will not permit successful recovery. In this connection the results of Spence and Domoney (1974) and Spence (in press) should be heeded. Their results suggest that the minimum adequate fraction, F, be calculated according to the following formula:

$$F = 6m/(n - 1),$$

where m is the number of dimensions in which the scaling is done, and n is the number of points. (See Spence, in press, for the rationale.) For example, if 30 points are to be scaled in three dimensions, at least 62% of the possible pairwise judgments should be collected. As n increases, this fraction becomes smaller; however, if high dimensional solutions are required, the necessary fraction will be larger.

In conclusion, we would like to make a comment on a piece of folklore familiar to all multidimensional scalers. This is the caution against interpreting, or attaching significance to, the relative positions of points that are close together in the recovered space. In Table 1, it is seen that in the no-deletion condition the mean absolute error does not vary much over the small, medium, and large distances. This means that *relative* error in the smaller distances is much greater than in the large distances. This effect is also seen in the cyclic-, small-, and medium-deletion conditions, although not in the large-deletion condition. We have tried to illustrate the consequences of this phenomenon in Figure 5. Assuming that the error in a recovered distance is a function of the uncertainty in locating the points, it seems reasonable, given the results in Table 1, to assume that this uncertainty is about equal for all points, thus producing a constant mean absolute error in the distances. Some possible recovered locations are indicated in Figure 5 by the open and solid circles for both a small and a large true distance. These are in the same relative positions in the regions of uncertainty. It is clear that one is on much surer ground when considering the relative location of points that are far apart.

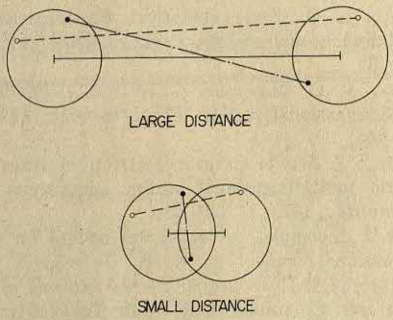

Figure 5. An illustration of the problem of attaching significance to the relative positions of points that are close together.

Reference Notes

1. Young, F. W. *A FORTRAN IV program for nonmetric multidimensional scaling* (L. L. Thurstone Psychometric Laboratory Report #56). University of North Carolina at Chapel Hill, 1968.
2. Johnson, R. M., & Van Dyk, G. J. *A resistance analogy for the efficiency of paired comparison designs.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, April, 1975.

References

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, *29*, 1–27.

Ramsay, J. O. Some statistical considerations in multidimensional scaling. *Psychometrika,* 1969, *34,* 167–182.

Ramsay, J. O. Maximum likelihood estimation in multidimensional scaling. *Psychometrika,* 1977, *42,* 241–166.

Spence, I. A Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms. *Psychometrika,* 1972, *37,* 461–486.

Spence, I. Incomplete experimental designs for multidimensional scaling. In R. B. Golledge & J. N. Rayner (Eds.), *Multidimensional analysis of large data sets.* Columbus: Ohio State University Press, in press.

Spence, I., & Domoney, D. W. Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika,* 1974, *39,* 469–490.

Tschudi, F. *The latent, the manifest, and the reconstructed in multivariate data reduction models.* Unpublished doctoral dissertation, University of Oslo, Oslo, Norway, 1972.

Wagenaar, W. A., & Padmos, P. Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *British Journal of Mathematical and Statistical Psychology,* 1971, *24,* 101–110.

Received September 8, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 67–79

Toward a General Model of Small Group Productivity

Samuel Shiflett
New York University

A general model of small group performance is proposed that is designed to encompass a wide variety of models of group performance extant in the psychological literature. These existing models are shown to be special cases of the more general model proposed here, and in fact, this model is developed from a foundation of these more restricted models. In developing the general model, distinctions are made between single and multiple resources and between unique and redundant resources. The relationship between problem-solving tasks and decision-making tasks is demonstrated by means of the model. Some examples of the model's usefulness for the understanding of group processes, leadership, and social decision schemes are presented.

Research on small group performance has been moving steadily toward a stance in which performance, actual or potential, is characterized by some sort of mathematical model involving various situational and personal characteristics as input parameters. One result of this theoretical trend has been a proliferation of fairly simple models that are rather restricted in their application because of various limiting assumptions frequently involved in the development of the models. The purpose of this article is to demonstrate that a number of existing models of group performance can be thought of as special cases of a single, more general model. The development of the model begins below with the presentation of several simple models extant in the literature and then proceeds to the more complex models, all the while demonstrating that each of these currently existing models can be treated as a special case of a single, more general model. In the process, several issues relating to the creation and use of mathematical models are touched on. The article concludes with a discussion and illustration of how a mathematical model of this sort can guide the development of so-called content-oriented theories. Finally, the possible application of the proposed model to such topics as leadership, organizational structure, and group decision making is illustrated.

To begin with, three general classes of variables are proposed: *resources*, *transformers* and *outputs*. Outputs are defined as any products that can be considered as an outcome of group interaction and include objective measures of group performance but might well include more subjective measures such as job satisfaction. Resources constitute "all the relevant knowledge, abilities, skills, or tools possessed by the individual(s) who is attempting to perform a task" (Steiner, 1966, p. 274). Resources are the raw materials that are essential for the creation of the product and without which the product could not exist. Transformers constitute all the variables that have an impact on resources and determine the manner in which they are incorporated into and related to the output variables. Transformers include such variables as situational and task constraints, role systems, and certain personal characteristics that may affect the way personal task-relevant resources are utilized in the output. An input variable can be a resource in some circumstances and a transformer in others, depending on the nature of the group product and the function or purpose of the resource. For example, oratorical ability acts as a resource when the group product is a theatrical piece, but it is a transformer when

The author thanks James H. Davis, Noel Dunivant, Raymond A. Katzell, and David Meister for their valuable comments on earlier versions of this article.

Requests for reprints should be sent to Samuel Shiflett, Psychology Department, New York University, 6 Washington Place, New York, New York 10003.

the thespiantics are used during group interaction to influence the manner in which resources are incorporated into a group product not directly requiring or reflecting any theatrical resources.

This particular categorization of variables was chosen because it was both simple and adequate for the development of the general model. The system is not even particularly original; it simply groups variables within a slightly different scheme than other investigators have suggested. For examples of slight variations on this scheme, see Hackman and Morris (1975), McGrath and Altman (1966), Naylor and Dickinson (1969), and Steiner (1972).

To summarize, input and output variables have been categorized in a manner that allows the model to be stated, in it simplest form, as $P = f(T, R)$, where P represents the group output or product, T stands for transformer variables, and R represents resource variables.

In 1966, Ivan Steiner published an important theoretical article in which he outlined five basic models of small group performance, relating his models to others already existing in the literature. He subsequently wrote a book (Steiner, 1972) on small group productivity, using that article as the organizing foundation for a more elaborate consideration of group processes. The first four of Steiner's models involve a single type of resource; the fifth model represents situations requiring more than one resource for the group product. The single-resource models are discussed together as a first step in the elaboration of the general model. The multiresource model is then considered as a somewhat more complex level of the general model.

Single-Resource Models

Steiner's (1966) *additive model* describes situations in which the task demands require every member of a group to perform the same function in a manner that causes the members' resources to enter the potential group product in a simple, additive manner. Under these conditions, average potential group productivity will vary as a positive, linear function of group size. The *disjunctive model* describes a situation in which the potential productivity of a group is determined entirely by the resources of its most competent member. The *conjunctive model* describes tasks in which potential productivity is restricted to a level that is established by the group's least competent member. The *compensatory model* describes a situation in which biases or errors in individual resources or products, such as judgments, are normally distributed and thus tend to cancel themselves or compensate among themselves and the group product is essentially an average of all the members' resources.

It is possible to demonstrate that all of Steiner's models can be treated as special cases of a more general model, which can be written as

$$P = \sum_{i=1}^{n} T_i R_i, \tag{1}$$

where P represents (actual) group productivity, R is the task-relevant resource of the ith person, and T is a weight representing the sum total of all constraints operating on the utilization of resources. In the case of Steiner's additive model for predicting potential productivity, T has the same value for all individuals in the group and can be written

$$\hat{P} = nT\bar{R}, \tag{2}$$

where P represents potential productivity, $\bar{R}$ represents the average ability of each group member, and n is the group size. It should be noted that the formula for *actual productivity* (P), strictly following Steiner's model, would be

$$P = \textit{potential productivity} - \textit{motivation and coordination losses.}$$

However, in the present case, motivation and coordination losses are considered to be a part of T acting on R in a multiplicative manner. Thus, T can be thought of as a function of task constraints, motivation losses, coordination losses, and in fact, any other variable that impinges on group resources in such a way as to change potential productivity from the ideal case expressed in Equation 1, when $T = 1$.

With the additive model now in hand, it is a simple matter to show that both the disjunctive and conjunctive models are also special cases of the general model, in which group members are thought of as constituting an

ordered set of individuals. To show this, Equation 1 can be rewritten to reflect the rank ordering of members from least capable to most capable, as follows:

$$P = T_L R_L + T_{L+1} R_{L+1} + \cdots + T_{M-1} R_{M-1} + T_M R_M. \quad (3)$$

Here, L represents the least capable member and M represents the most capable member. When task and situational constraints are such that only the best member's performance determines productivity, that is, if task constraints are disjunctive, then $T_M > 0$, all other $T = 0$, and the disjunctive model is characterized by the equation,

$$P = T_M R_M. \quad (4)$$

Note that T_M is not necessarily equal to 1, since other constraints can still moderate the extent to which resources become products. However, T_M is the only transformation weight greater than 0. As was pointed out by Steiner (1966) and Davis (1969), this model is essentially the same as Lorge and Solomon's (1955) "Model A," Thomas and Fink's (1961) "rational model," and the model proposed by Taylor (1955). Also closely related is Steiner and Rajaratnam's (1961) method for predicting group competence levels.

In a similar fashion, when task and other situational constraints are conjunctive in nature, $T_L > 0$, and all other $T = 0$, resulting in the conjunctive model,

$$P = T_L R_L. \quad (5)$$

The compensatory model can similarly be expressed in terms of the general Equation 1 if T_i takes the value C_i/n, where n is the group size and C_i represents the impact of all other constraining variables, including the distribution of errors. In fact, any averaging model is a transformation of the general additive model in which T has been appropriately adjusted for the group size.

Steiner (1972) subsequently introduced the concept of a *discretionary task* to describe conditions in which the group members can combine their individual resources in any manner they wish. This is another way of saying that all four models described above are special cases of the general model except that, for Steiner, a discretionary task implies that group members themselves are able to determine what values T takes instead of the values being imposed by situational characteristics not under group control. Similar points have been made by Hackman and Morris (1975) and Shiflett (1972).

Multiple-Resource Models

Steiner's (1966) final category of task types, the complementary tasks, are designed to deal with cases in which a single individual performs only a part of the total task while other group members, possessing different kinds of resources, perform the remaining parts. Steiner further subdivided complementary tasks into those with *unshared resources* and those with *partially shared resources*. Steiner originally argued that to be complementary the task must be divisible into subtasks and that the resources required for the various subtasks must differ in a qualitative manner. Subsequently, Steiner (1972) emphasized the divisibility requirement by referring to these types of tasks as *divisible tasks* and stressed the idea of appropriately matching resources to the requirements of the various subtasks. Laughlin and Branch (1972) suggested that although this strategy can be referred to as a division of labor, it can also be reasonably seen as a division of resources. Thus, in order to optimize or maximize group performance, the division of resources must be coordinated with the partitioning of the task, whether accomplished by the group itself or by a superordinate organization.

Under these conditions the distinction between shared and unshared resources disappears, since by appropriate assignment of individuals to subtasks, resources may be redistributed so that they are not shared on any subtasks even though they are shared within the group as a whole. For this reason, the unshared resources model can be treated as a special case of the more general shared resources model. Similarly, it is possible to demonstrate that all the single-resource models are special cases of the unshared resources model, which, as we have just shown, is a special case of the shared resources model.

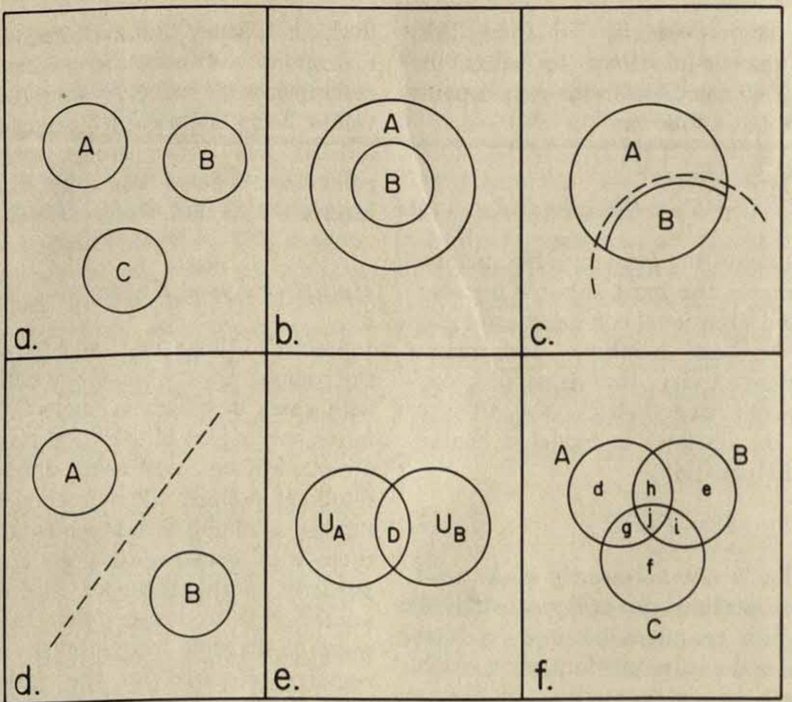

Figure 1. Venn diagrams illustrating various distributions of resources within a group.

To illustrate these propositions, set theory operations can be utilized to describe a series of identities. The general model, expressed in Equation 1, can now be redefined as the union ($\bigcup$) of all members' resources available for the group product:

$$P = T_1R_1 \bigcup T_2R_2 \bigcup \cdots \bigcup T_{n-1}R_{n-1} \bigcup T_nR_n. \quad (6)$$

T_i acts as a constraint on R_i, changing it from the total amount potentially available from the group member into the amount actually available as a result of various task and situational conditions. Proceeding as before, we can now redefine the four basic models already described.

The additive case exists when

$$T_1R_1 \bigcup T_2R_2 \bigcup \cdots \bigcup T_nR_n = R_1R_1 + T_2R_2 + \cdots + T_nR_n = \sum_{i=1}^{n} T_iR_i. \quad (7)$$

This situation can be illustrated by the use of Venn diagrams, as in Figure 1a, and is the case defined by Equation 1. That is, in terms of Figure 1a, the additive case exists when $A \bigcup B \bigcup C = A + B + C$.

The disjunctive case exists when

$$T_1R_1 \bigcup T_2R_2 \bigcup \cdots \bigcup T_nR_n = T_MR_M. \quad (8)$$

Here, $T_M > 0$, and all other $T = 0$. The conjunctive case exists when

$$T_1R_1 \bigcup T_2R_2 \bigcup \cdots \bigcup T_nR_n = T_LR_L. \quad (9)$$

In this case, $T_L > 0$, and all other $T = 0$.

It should be evident that in keeping with the assertion that all three models are special cases of the more general model, the left side of the three expressions does not change, for it *is* the general model.

The disjunctive and conjunctive cases can occur in several different ways, which are illustrated in Figures 1b, 1c, and 1d. In Figure 1b, the resources of the least able member (B) are completely encompassed by those of the

most able member (A). This overlapping of resources is sometimes referred to as redundancy. In this situation, if no other constraints exist, then Member B is unable to add anything more to the product than what Member A already has the potential to contribute. In other words, the disjunctive case exists and $A \cup B = A$. On the other hand, if some constraint exists so that the appropriate combination rule is $A \cup B = B$, then the conjunctive case exists. This situation is illustrated in Figure 1c, where the dashed line symbolizes a constraint preventing the use of A's superior resources.

In Figure 1d, resources are not shared, in contrast with the shared resources situation of Figures 1b and 1c. The dashed line again represents external constraints that operate in a manner that prevents one member from utilizing his or her resources or in which productivity is totally dependent on one member. If productivity is determined by the most able member, then this particular situation is disjunctive, but if productivity is totally dependent on the least able member, then the situation is conjunctive. It should also be evident that the additive model (Equations 1 and 7 and Figure 1a) is equivalent to the unshared resources model, whether or not the resources are qualitatively similar.

Compensatory situations can exist in a variety of ways, but it can be said that a compensatory model is appropriate when the various constraints on resources operate in such a way that the following identity is true:

$$R_1 \cup R_2 \cup \cdots \cup R_{n-1} \cup R_n = \left(\sum_{i=1}^{n} R_i\right)/n. \quad (10)$$

The Model in Terms of Matrix Operations

Resources within a group can be thought of as being arrayed in a resource matrix, **R**, in which rows are defined by group members (m), and columns are defined by a discrete, mutually exclusive form of resource (a),

$$\mathbf{R} = \begin{array}{c} \\ m_1 \\ m_2 \\ \\ m_i \end{array} \!\!\begin{array}{c} \begin{array}{cccc} a_1 & a_2 & \cdots & a_j \end{array} \\ \begin{bmatrix} r_{11} & r_{12} & & r_{1n} \\ r_{21} & & & \\ & & & \\ r_{m1} & r_{m2} & & r_{mn} \end{bmatrix} \end{array}. \quad (11)$$

Entries consist of the amount of each resource (r) provided by each member. In a single-resource situation the matrix would be a column vector. In all cases, resources are assumed to be directly relevant and necessary for creating the group product, whatever it is defined to be. Thus the nature of this matrix will vary widely depending on the nature of the group product. Often, each column will represent a qualitatively different resource.

These different resources must be combined in some fashion to finally represent a group product. This combination rule usually involves rescaling the resources for proper reflection in the product, but it also requires specification of a particular rule. Although this task is formidable, several approaches to the problem have been suggested. Job analysis (Zedeck & Blood, 1974) and synthetic validity (Guion, 1965) represent approaches to breaking products (or their antecedent tasks) into the subsets of skills and resources required to accomplish the task. The entries can vary in nature from interval measures of skills and abilities to probabilities ranging from 0 to 1 of a certain response being in an individual's behavioral repertoire.

Implicitly or explicitly, all theories of group performance specify one or more rules for the manipulation of the resource matrix. The rules are specified in a transformation matrix, **T**. In the additive model, **T** is a vector in which all entries take on the same value; **R** is also a single vector, indicating the amount of the resource provided by each member. The conjunctive and disjunctive models specify a **T** vector containing all entries equal to 0 except for the one entry specific to either the most or least able member.

As the product and the model become more complex, so do the matrices representing the model. In particular, the transformation matrix may be the final representation of a whole series of matrix operations representing some model of group process. An example of this explicit use of matrix operations to represent group processes is structural role theory (Oeser & Harary, 1962, 1964; Oeser & O'Brien, 1967). The basic elements in the model consist of persons, positions, tasks, and the various directional relationships among these elements. These elements are then com-

bined in a series of mathematical operations that result in a specific index describing some aspect of the situation. The particular operations are a function of the index of interest, but the elements are always those listed above. Elaborating on the basic model, O'Brien and his associates have developed indexes of collaboration requirements, cooperation requirements, certain task characteristics, and potential leader influence (O'Brien, 1968, 1969a, 1969b; O'Brien, Biglan, & Penna, 1972; O'Brien & Owens, 1969; Witz & O'Brien, 1971). Basically, then, structural role theory can be thought of as one approach to specifying the transformation matrix, **T**.

Unique Versus Redundant Resources

We have already noted the fact that resources can be distributed within a group and applied to the group product in a variety of ways. Continuing the use of set notation, resources are characterized below as either duplicated (redundant) or unique. If two individuals' resources are characterized by the Venn diagram illustrated in Figure 1e, then the following sets can be defined: A = Member A's total resources; B = Member B's total resources; U_A = A's unique resources, those not possessed by B; U_B = B's unique resources; and D = redundant or overlapping resources possessed by both A and B. From these definitions the following sets can be further identified: $A \supseteq (U_A \cup D)$; $B \supseteq (U_B \cup D)$; $U \supseteq (U_A \cup U_B)$.

As a matter of fact, a total set of resources can be characterized in any number of ways, depending on the requirements of the particular model or theory being espoused. For example, some models of performance have emphasized the amount of unique resources available, sometimes referred to as the pooling of resources, as in the case of Kelley and Thibaut (1969) and that of the research by Laughlin and his associates (Laughlin & Branch, 1972; Laughlin, Branch, & Johnson, 1969; Laughlin & Johnson, 1966). Other models have emphasized the impact of the amount of redundancy, or overlap, on performance, for example, Zajonc and Smoke (1959) and Shiflett (1972, 1973, 1976).

The number of sets of overlapping data increases rapidly as group size increases. In groups of size two, there is one set of shared resources. In groups of size three, there are four sets of overlapping resources, including the set that is common to all three members. This situation is illustrated in Figure 1f, and the sets are labeled g, h, i, and j. When $N = 4$, *common resource sets* = 11, and when $N = 5$, *common resource sets* = 25. Obviously, this approach to the subdivision of resources rapidly becomes unwieldy as group size increases, but it does illustrate the fact that the set of available resources can be broken into subsets, using any number of different criteria. And though the number of sets can quickly burgeon into an unmanageable size as group size increases, in practice this is unlikely to represent a major problem, since simplifying procedures and assumptions can be applied that substantially reduce the number of sets that would have to be dealt with. For example, it seems likely that many of the subsets of shared resources will be trivially small or nearly identical to other subsets, so resources could be regrouped into more manageable and possibly more meaningful sets.

The distinction between unique resources and shared resources provides several interesting implications for the manner in which individual resources will be utilized in the group product when the resources consist of sets of discrete units of the resource and when the task is subdivisible. For example, it should now be apparent that the fact that two or more individuals possess the same resource does not increase the total set of available resources but does increase the probability of that resource being used. Recognizing that some resources are shared provides one explanation for the frequent finding that groups are less productive than would be expected by a knowledge of each individual's resource score. In other words, this distinction permits the recognition of the fact that $(A \cup B) \leq (A + B)$.

Although the two sets of variables, unique and redundant, are potentially independent of each other, in most cases they probably are not. They must always be considered together in looking at group performance, even if one or the other turns out to be an empty set.

Failure to consider both sets adequately can lead to inadvertent confounding of one with the other. For example, in the series reported by Laughlin and his associates (Laughlin & Branch, 1972; Laughlin et al., 1969; Laughlin & Johnson, 1966), subjects of varying degrees of ability (resources) were composed into 2-, 3-, or 4-person groups in factorial combinations of member ability. The basic purpose of the studies was to test Steiner's (1966) complementary model, which in this context asserts that groups with members having nonredundant (unique) resources will do better than groups with all members having the same (redundant) resources, assuming that all resources are relevant to the task product.

The presence and amount of unique resources in a group were inferred by Laughlin and his associates by looking at the difference in ability levels of the group members. Thus, a member of low ability was predicted to be unable to contribute any new resources to those of a member of high ability, and productivity would be determined entirely by the most able member's resources, that is, using the present notation, $P = U_M + D$. In the case of equal ability, however, it was assumed that both provided some unique resources (i.e., $U_A > 0$ and $U_B > 0$) in addition to a core of overlapping resources, D. Thus, in this particular case, P was assumed to equal $U_A + U_B + D$ and was predicted to be greater than when either member worked alone or with a member of lesser ability.

The data appeared to support these predictions strongly. However, there are several problems in the design and interpretation of these studies. In the case of members of unequal ability, we notice that although the complementary model does indeed imply the prediction that $P = U_M + D$, so does the simpler disjunctive model. As was illustrated at the beginning of this section, any member's abilities can be thought of as containing the subsets U and D. Thus the most able member's abilities, R_M, are composed of unique resources, U_M, as well as those shared by other members, D, or $R_M = U_M + D$. By simple substitution we see that the prediction of the complementary model, $P = U_M + D$, is identical to that of the disjunctive model, $P = R_M$.

A more important problem in the Laughlin series consists in the definition of unique and redundant resources. The amount of unique or pooled resources is clearly varying systematically as a function of variations in the ability level of group members, as was asserted by the researchers; however, the amount of redundancy is also varying systematically, and the precise relationship of one set to the other cannot be determined from the data. Thus, it is not possible to state unequivocally that uniqueness of resources determined the results, since redundancy could equally have been the cause. Or more likely, some combination of the two sets explains the results. In order to partition these effects, empirical measurements of both redundancy and uniqueness of resources must be made.

The difficulty in assessing redundancy in most small group research can be illustrated by looking at two individuals who each score 50 on a 100-item vocabulary test. How much redundancy is represented in the two scores of 50? If each person answered 50 items correctly we could argue that there is 100% redundancy. However, this condition would hold only if both members correctly answered exactly the same items. If all the items are of equal difficulty, then redundancy could range from 0% to 100% in the case of 50 of 100 items. It should be obvious, then, that it is not possible to infer directly the amount of redundancy from ability levels alone, although we would expect that as one member's ability becomes increasingly greater than another's, $A \cap B \rightarrow A$.

Much of the decision-making research avoids this problem because it is possible to precisely specify the redundancy term by causing it to be a characteristic of the environment (the **T** matrix) rather than the resource matrix (Slovic & Lichtenstein, 1971). Shiflett (1972, 1973, 1976) was also able to manipulate redundancy by arbitrarily dividing the task in several ways, thus making redundancy a function of **T** instead of **R**. However, he was not any more able to estimate the actual amount of redundancy present in the shared-labor work strategy than were Laughlin and his associates in their shared-labor task.

The general model, written to reflect the distinction between shared and unique re-

sources, is

$$P = \sum_{i=1}^{n} T_i U_i + \sum_{i=1}^{n} T_i^* D, \qquad (12)$$

where U_i represents the unique resources of Individual i, T_i is the transform weight reflecting the utilization of those resources in the group product P, and T_i^* is the transform weight applied to redundant resources, D. T and T^* may differ depending on the particular constraints operating in a group. It might seem that T^* should be constant across all individuals, since D is defined to be, but again, particular constraints operating in a group situation may cause this not to be so. That is, some people's common resources are more likely to be used than are those of others.

Discussion

To illustrate how the proposed model might be used, two examples of translation from mathematical abstractions to psychological realities are presented below. The first illustration comes from the area of leadership, the second considers some current approaches to decision making.

Leadership

In terms of the present model, leadership may be defined as a *resource recognition* or a T-*facilitating* function. That is, leaders must recognize the existence of resources available to them through their group members, and they must be able to do so to the extent that they can locate those resources with enough precision that they can then take appropriate actions to facilitate their effective inclusion in the group product. Facilitative behaviors may include a wide variety of actions, from restructuring an organization in order to better utilize the existing distribution of resources to soothing angry feelings and increasing the motivation and morale of the group so that the members are willing to apply their resources to the task. Thus a leader's appropriate or most effective behavior can vary drastically, depending on the situational requirements, but in every case the behavior should have the same end result: to maximize P by making appropriate adjustments to the T weights over which she or he has some conscious control.

It would seem that the approach to leadership suggested here fits rather nicely with the ideas recently proposed by Graen and his associates (Dansereau, Cashman, & Graen, 1973; Graen, 1976). They suggested that the appropriate unit of analysis for examining leadership processes is the "vertical dyad," consisting of the superior and only one subordinate of the several in a work unit. In this approach the vertical dyad reflects the role relationship linking each group member and his or her superior. To the extent that this process link affects an individual's contribution to the group product it is represented in the model by T. Each vertical dyad can be represented by a single number reflecting "influence," and the set of all vertical dyads in a work unit can then be written as a vector **T**, containing a series of *t*s, one for each group member or vertical dyad. It will be recalled that we pointed out that the **T** matrix is the culmination of what might be a whole series of preliminary matrix operations reflecting various aspects of the situation. What has traditionally been called *leadership style* is then simply a description of the various components of the summary transformation index that the leader actually influences in his or her attempts to increase performance.

The effectiveness of any leadership style is a function of the context within which the style is exercised. In some situations it would be more appropriate to emphasize organizational factors, whereas in others it would be more appropriate to influence motivational factors. Obviously, a good leader is one who emphasizes the right components at the right time. A flexible leader is one who can change emphasis or style as situational requirements change. This approach also allows us to see a common ground uniting Graen's approach with Likert's (1958) idea that leadership is an adaptive process and Naylor and Dickinson's (1969) idea that a group's communication structure, which includes leader communication, is dependent on task structure and work structure.

Vroom and Yetton (1973) proposed a model of leadership as a decision-making process that also appears to be related to the approach being

proposed here. They suggested that a leader has a series of discrete, alternative behavioral processes that can be used in a problem-solving situation. The alternatives can range from behaviors that could be labeled "autocratic" through those called "participative." The leader must make a series of sequential decisions in evaluating each situation in order to determine the most appropriate behavior for that situation. Thus, their model stresses the need for behavioral differences on the part of the leader over situations. This same implication is a keynote of the vertical-dyad approach of Graen and his colleagues and is a direct implication of the present model.

It is noteworthy that the effect of a participative style of leadership is effectively to assign an equal weight to all members' resources, whereas the autocratic style is characterized at its extreme as one with a matrix of group members' weights equal to 0 and a leader weight equal to 1.

Vroom and his associates noted that maximizing group efficiency or productivity is not always the only criterion for determining a behavioral style. Other considerations, such as humanistic concerns or quality of interaction or a pragmatic desire to develop certain social characteristics of the members might also influence the particular leader style. The latter decision base was termed Model B by Vroom and Yetton (1973), and the decision model involving only group productivity was termed Model A. Using the rationale of the present model it appears that Model A is in fact a subset or special case of the more general Model B. By defining group productivity, P, as the set of all possible outcomes and products resulting from group processes, then P is a set containing not only traditionally defined group performance criteria but also group-level perceptions such as morale or group atmosphere and individual-level perceptions, attitudes, and characteristics such as job satisfaction and feelings of acceptance, worthwhile accomplishment, or personal growth. To each of these elements in the product set a weight reflecting something like importance or relevance can be attached forming a vector, $\mathbf{I}$, with elements equal in number to those contained in the $\mathbf{P}$ vector. These importance or relevance weights contained in the $\mathbf{I}$ vector are frequently, but not always, determined by some authority external to the group or by the leader of the group. Thus the president of a production firm may determine that the only relevant product from an assembly line is the physical product being assembled and nothing else. In this case, a relevance weight of 1 is attached to the group product and weights of 0 are attached to all other nonmaterial group outcomes, resulting in an extreme example of Model A. Obviously, there is no reasonable case in which all potential outcomes can be simultaneously considered, but any case in which one or more nonmaterial outcomes are given weights greater than 0 can be considered to be a variation of Model B. Variants of Model B can be defined by the appropriate importance vector. Thus sensitivity, therapy, or team-building groups would usually be characterized by a vector containing weights of 0 for material group products but large weights for personal or interpersonal characteristics deemed important to the group.

Decision-Making Tasks

In a previous section it was suggested that one useful way of conceptualizing the resource matrix was in terms of its redundancy. A summary vector of redundancy could consist of a row of entries, each of which indicates how many members have a particular resource. Thus the vector would contain as many entries as there are defined resources. Using this concept it is possible to demonstrate a close relationship between problem-solving or general performance tasks and decision-making tasks. I am suggesting that many decision tasks require as their basic matrix the redundancy vector, whereas problem-solving tasks require the summary resource matrix or some combination of the resource and redundancy matrices. The reason for this distinction is found in the distinctly different way in which resources are incorporated into the group product in the two types of tasks. Resources (or response alternatives) in a decision task are mutually exclusive, and the use of one response alternative precludes the use of the remaining response alternatives. That is, if my response alternatives are *yes* and

no, I cannot use both responses at the same time. If I try to, I really have not made a decision. In a problem-solving task, on the other hand, resources can be combinatorial so that the use of one resource does not preclude the use of additional resources but actually supplements or combines with other resources to reflect the final product.

Pursuing the special characteristics of decision tasks, we find that there is typically a clearly defined set of response alternatives, which is frequently quite limited in scope, often consisting of only two or three alternatives. Because of this fact and because the response alternatives are mutually exclusive, it is convenient to think of each response alternative as a qualitatively distinct resource. Thus, a two-alternative decision task is defined as requiring two resources, a three-alternative decision task has three resources, and so forth. Under these circumstances, the concept of redundancy takes on special importance. By definition, each member has only one resource (response alternative). Thus it is not possible for subjects to have multiple resources. Because response alternatives are always mutually exclusive within a group member, the **R** matrix has the characteristic that each subject vector contains entries consisting of $r - 1$ 0s and a single 1, for example,

$$\mathbf{R} = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \end{array} \begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}.$$

This characteristic simply is reflecting the fact that when one response is made, the other cannot occur. The redundancy matrix, **D**, is then the simple sum of resources across subjects; using the same example as above,

$$\mathbf{D} = \begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_3 \end{array} \\ \begin{bmatrix} 2 & 1 & 1 \end{bmatrix} \end{array}.$$

Because of the way decisions are typically made, it is the redundancy vector that is the primary resource matrix of interest in this type of task. Transformation rules (**T** matrix) thus operate on the **D** matrix of redundancy rather than on the original resource matrix, **R**.

Recently, Davis (1973), in an important theoretical article, proposed a model for describing the nature of the social processes leading to a group decision that is a perfect illustration of this conceptualization of resources. His theory assumes that, as suggested above, response alternatives defined by the decision task are mutually exclusive and exhaustive. An individual probability distribution is assumed to exist that characterizes the population from which a decision-making group is formed. Randomly formed groups from such a population display an internal distribution of preferences that Davis called a "distinguishable distribution" and that is identical to the redundancy vector in the present model. To quote the succinct summary statement by Davis, Kerr, Sussman, and Rissman (1974), "In this system, response alternatives, but not particular people, are considered distinguishable" (p. 250). In the Davis model the population's individual probability distribution (the population redundancy vector) is operated on by a "social decision scheme" matrix to yield an estimate of the probability of a group deciding for each alternative. The social decision scheme matrix is a mathematical statement describing how each possible group-preference distribution is resolved through social interaction, and as such, is equivalent to the presently defined transformation matrix, **T**.

Davis's (1973) approach can be seen as a special treatment of unique and redundant resources in which the social decision scheme determines or influences the relative importance of the unique and redundant resources. In this scheme, a shared resource is any resource that has more than one adherent in the group, and a unique resource is one that has only one proponent. Thus, with respect to the set of response alternatives, $A_1, A_2 \cdots A_n$, and occupancy numbers (redundancy), $r_1, r_2 \cdots r_n$, it can be said that when $r_n = 1$ for any A_n, that set is a nonredundant or unique set. And when $r_n > 1$ for any A_n, then the set is a redundant set.

The impact of redundancy on the final outcome will be determined by the social decision scheme operating on the resources. Some of the social decision schemes have been given psychologically descriptive titles by Davis, examples of which are "majority, equiprobability, plurality, proportionality, and

truth-wins." These matrices can represent very formally presented rules that are agreed on by the group, or more subtle and unarticulated influences such as cultural mores, or even task characteristics.

Conclusion

A number of factors have been shown to influence the relative contribution of resources (R) to the group product (P). In addition, many factors potentially enter into the system of formulas for estimating T. In fact, the system is usually so overdetermined that unique solutions seldom exist. Any number of different combinations of group-task or transformational factors can result in essentially the same **T** matrix. Thus a host of influences might affect a group, both internally and externally, and yet the group can still make adjustments so that little change in actual productivity occurs. However, many important constraints on the resource–product relationship are often not amenable to change. Under these conditions many factors used for estimating T have the status of constants and the system rapidly moves from its overdetermined status to one in which there are few solutions. Any attempts to improve or change group performance by affecting the **T** matrix are severely handicapped under these conditions. A simple example illustrating this situation would be a routinized assembly line in which all operations are highly determined and possible variations in operating procedures are severely limited.

Of course, in some cases neither the **T** nor the **R** matrix can be substantially changed without incurring great costs. The assembly line again serves as an example. Once workers have been adequately trained, the resource matrix has reached a maximum—There is no way to increase resources within the constraints of the assembly line. To the extent that changing the assembly line itself (i.e., the **T** matrix) involves extensive retooling of expensive machinery, neither matrix is likely to be changed unless other considerations that are unrelated to the **R** or **T** matrix, such as social or humanitarian concerns, are brought to bear on the decision to change the situation.

The extent to which the present model will adequately characterize a group's resource–transform–product system is partially dependent on the time frame within which the model is to characterize the group. In general, the longer the time interval, the more opportunity there is for unexpected perturbations in the environment to occur and thus for the model to be inadequate. For example, the amount and distribution of resources within the group do not have to remain constant. Thus the amount of resources available within the group can change drastically over a very short period simply as a result of the transformation processes within the group. A group can start with each member bringing unique resources to the task at hand; that is, task-relevant redundancy equals 0. But as a result of group interaction, not only has a product been created that presumably reflects the various resources but the group members may have distributed duplicate resource sets among themselves. Task-relevant redundancy now approaches 100% as a result of learning. When this occurs, the resource matrix changes substantially, and this in turn may affect the adequacy of the existing **T** matrix.

As the group interacts over time, interim products occur both in terms of the current status of the formal group product and in terms of interpersonal relations and individual perceptions. These interim outputs are frequently continuous in nature, but by arbitrarily dividing the group interval into small time units, a series of interim products can be identified as the group progresses toward its final goal. These interim product states act as inputs to the next time frame, influencing the manner in which member resources can be incorporated into the group product. If, for example, performance on a subtask changes the nature of the resources available for the remaining portion of a task, the final group product can also be expected to differ from what might be expected from a knowledge of the resources available at the beginning of the task interval. The Lorge and Solomon (1955) Model B is an early attempt to incorporate this fact of life into a mathematical model. Restle and Davis (1962) and Davis and Restle (1963) inferred group-performance characteristics over time as a function of the number

of distinct, sequential subtasks within a group task. Recently, Davis (1973) presented an interesting analysis of the potential effect of sequential decision making on interim resource matrices as a decision-making task moves up an organizational hierarchy.

The model proposed here attempts to demonstrate that a number of existing models of group performance can be seen as existing within a single, unified conceptual framework that still retains a fair degree of simplicity. In fact, given the present framework, it is possible to see that most of the models in the literature are really only special cases of this model, in which a specific pattern of T weights is postulated to have a specific effect on group resources (R). This approach has often taken the form of comparing two or more T patterns (e.g., conjunctive vs. disjunctive, or equiprobability vs. majority rule vs. truth-wins). In some cases the **R** matrix is the primary focus. Only rarely does one come across careful consideration of both the **T** matrix and the **R** matrix in the same study. Unfortunately, these cases are restricted almost entirely to the decision-making literature (e.g. Davis, 1973; Slovic & Lichtenstein, 1971).

It should be kept in mind that the present model can be thought of as a representational framework, incorporating a variety of smaller models. As such, in its current form it has only limited capability for generating precisely formulated predictions. In other words, the model does not represent a new theory of group productivity, nor does it solve the very real problems of assessing resource and transformation variables or of determining the best fit between the **T** and **R** matrices. Research along all of these fronts is in progress and appears to be making slow but steady advances. Hackman and Morris (1975) presented an excellent review and discussion of many of these practical problems. I hope that the approach presented here, by illustrating the need to consider the transformation and resource variables simultaneously and by illustrating the underlying similarities among a variety of existing models, will provide a general framework within which these other issues can be more clearly defined and resolved.

Research incorporating both transformation and resource variables is growing and has, I believe, an exciting future. Questions about, for example, the ideal T weights for a given resource set are now being asked, as well as broader questions regarding the effects of various patterns of T weights where characteristics of the **R** matrix are not entirely known. Since ideal weighting is often impossible because of lack of information or other systematic distortions in weighting due to group processes (Steiner, 1972), research is necessary to determine ways of minimizing bias or to determine the effects of various conditions on probable weightings. One method of attacking this problem has been to compare the effectiveness of various formally defined weighting schemes on group judgments. The research program of Davis (1973) and his associates illustrates this approach. Other recent research has demonstrated that an averaging or equal-weighting model outpredicts differential-weighting models in a variety of conditions and may be especially effective when the **R** matrix is not known (Einhorn, Hogarth, & Klempner, 1977). A different but related approach would be to determine the optimal distribution of resources for a fixed **T** matrix. The approach would be applicable to situations in which there is little freedom to vary the surrounding environment.

All of these approaches, with their different methods, terms, and focuses, are attempts to capture a general process by which individual resources are transformed into a group product through group processes. The present framework represents an early step toward bringing together these diverse models into a unified approach to the study of small groups, by concentrating on the similarities among these models. It is hoped that in doing so, future research will begin to come together into more meaningful patterns and that the current state of affairs in the study of group dynamics, in which theoretical integration is almost nonexistent (Shaw, 1976), will begin to move toward a state of genuine integration.

References

Dansereau, F., Jr., Cashman, J., & Graen, G. Instrumentality theory and equity theory as complementary approaches in predicting the relationship of leadership and turnover among managers. *Organizational Behavior and Human Performance*, 1973, *10*, 184–200.

Davis, J. H. *Group performance*. Reading, Mass.: Addison-Wesley, 1969.

Davis, J. H. Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, 1973, *80*, 97–125.

Davis, J. H., Kerr, N., Sussmann, M., & Rissman, A. K. Social decision schemes under risk. *Journal of Personality and Social Psychology*, 1974, *30*, 248–271.

Davis, J. H., & Restle, F. The analysis of problems and prediction of group problem solving. *Journal of Abnormal and Social Psychology*, 1963, *66*, 103–116.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. Quality of group judgment. *Psychological Bulletin*, 1977, *84*, 158–172.

Graen, G. Role making processes within complex organizations. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 1976.

Guion, R. M. *Personnel testing*. New York: McGraw–Hill, 1965.

Hackman, J. R., & Morris, C. G. Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 9). New York: Academic Press, 1975.

Kelley, H. H., & Thibaut, J. W. Group problem solving. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 4, 2nd ed.). Reading, Mass.: Addison–Wesley, 1969.

Laughlin, P. R., & Branch, L. G. Individual versus tetradic performance on a complementary task as a function of initial ability level. *Organizational Behavior and Human Performance*, 1972, *8*, 201–216.

Laughlin, P. R., Branch, L. G., & Johnson, H. H. Individual versus triadic performance on a unidimensional complementary task as a function of initial ability level. *Journal of Personality and Social Psychology*, 1969, *12*, 144–150.

Laughlin, P. R., & Johnson, H. H. Group and individual performance on a complementary task as a function of initial ability level. *Journal of Experimental Social Psychology*, 1966, *2*, 407–414.

Likert, R. Effective supervision: An adaptive and relative process. *Personnel Psychology*, 1958, *11*, 317–352.

Lorge, I., & Solomon, H. Two models of group behavior in the solution of Eureka-type problems. *Psychometrika*, 1955, *20*, 139–148.

McGrath, J. E., & Altman, I. *Small group research: A synthesis and critique of the field*. New York: Holt, Rinehart & Winston, 1966.

Naylor, J. C., & Dickinson, T. L. Task structure, work structure, and team performance. *Journal of Applied Psychology*, 1969, *53*, 167–177.

O'Brien, G. The measurement of cooperation. *Organizational Behavior and Human Performance*, 1968, *3*, 427–439.

O'Brien, G. E. Group structure and the measurement of potential leader influence. *Australian Journal of Psychology*, 1969, *21*, 277–289. (a)

O'Brien, G. E. Leadership in organizational settings. *Journal of Applied Behavioral Science*, 1969, *5*, 45–63. (b)

O'Brien, G. E., Biglan, A., & Penna, J. Measurement of the distribution of potential influence and participation in groups and organizations. *Journal of Applied Psychology*, 1972, *56*, 11–18.

O'Brien, G. E., & Owens, A. G. Effects of organizational structure on correlations between member abilities and group productivity. *Journal of Applied Psychology*, 1969, *53*, 525–530.

Oeser, O. A., & Harary, F. A. A mathematical model for structural role theory: I. *Human Relations*, 1962, *15*, 89–109.

Oeser, O. A., & Harary, F. A. A mathematical model for structural role theory: II. *Human Relations*, 1964, *17*, 3–17.

Oeser, O. A., & O'Brien, G. A mathematical model for structural role theory: III. *Human Relations*, 1967, *20*, 83–97.

Restle, F., & Davis, J. H. Success and speed of problem solving by individuals and groups. *Psychological Review*, 1962, *69*, 520–536.

Shaw, M. E., *Group dynamics: The psychology of small group behavior*. New York: McGraw–Hill, 1976.

Shiflett, S. Dyadic performance on two tasks as a function of task difficulty, work strategy, and member ability. *Journal of Applied Psychology*, 1976, *61*, 455–462.

Shiflett, S. C. Group performance as a function of task difficulty and organizational interdependence. *Organizational Behavior and Human Performance*, 1972, *7*, 442–456.

Shiflett, S. C. Performance effectiveness and efficiency under different dyadic work strategies. *Journal of Applied Psychology*, 1973, *57*, 257–263.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processes in judgment. *Organizational Behavior and Human Performance*, 1971, *6*, 649–744.

Steiner, I. D. Models for inferring relationships between group size and potential group productivity. *Behavioral Science*, 1966, *11*, 273–283.

Steiner, I. D. *Group process and productivity*. New York: Academic Press, 1972.

Steiner, I. D., & Rajaratnam, N. A model for the comparison of individual and group performance scores. *Behavioral Science*, 1961, *6*, 142–147.

Taylor, D. W. Problem solving by groups. In *Proceedings of the XIV International Congress of Psychology, 1954*. Amsterdam: North-Holland, 1955.

Thomas, E. J., & Fink, C. F. Models of group problem solving. *Journal of Abnormal and Social Psychology*, 1961, *63*, 53–63.

Vroom, V. H., & Yetton, P. W. *Leadership and decision-making*. Pittsburgh, Pa.: University of Pittsburgh Press, 1973.

Witz, K., & O'Brien, G. E. Coordination indices in work organizations. *Journal of Mathematical Psychology*, 1971, *8*, 44–57.

Zajonc, R. B., & Smoke, W. H. Redundancy in task assignments and group performance. *Psychometrika*, 1959, *24*, 361–369.

Zedeck, S., & Blood, M. R. *Foundations of behavioral science research in organizations*. Monterey, Calif.: Brooks/Cole, 1974.

Received September 16, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 80–92

Comparative Effectiveness of Paraprofessional and Professional Helpers

Joseph A. Durlak
Southern Illinois University at Carbondale

Forty-two studies comparing the effectiveness of professional and paraprofessional helpers are reviewed with respect to outcome and adequacy of design. Although studies have been limited to examining helpers functioning in narrowly defined clinical roles with specific client populations, findings have been consistent and provocative. Paraprofessionals achieve clinical outcomes equal to or significantly better than those obtained by professionals. In terms of measurable outcome, professionals may not possess demonstrably superior clinical skills when compared with paraprofessionals. Moreover, professional mental health education, training, and experience do not appear to be necessary prerequisites for an effective helping person. The strongest support for paraprofessionals has come from programs directed at the modification of college students' and adults' specific target problems and, to a lesser extent, from group and individual therapy programs for non-middle-class adults. Unfortunately, there is little information on the factors that account for paraprofessionals' effectiveness. Future studies need to define, isolate, and evaluate the primary treatment ingredients of paraprofessional helping programs in an attempt to determine the nature of the paraprofessional's therapeutic influence.

Evaluations of research involving paraprofessional therapists have been highly positive. Reviewers concentrating on the use of college students (Gruver, 1971), community volunteers (Siegel, 1973), and parents (Berkowitz & Graziano, 1972; Johnson & Katz, 1973; O'Dell, 1974) or discussing paraprofessional therapy outcome research in general (Karlsruher, 1974) have agreed that relatively untrained workers can make a significant service contribution within the mental health field. Conspicuously ignored, however, has been the research comparing the relative effectiveness of paraprofessionals and professionals. This article attempts to review and evaluate this comparative research with respect to outcome and adequacy of design.

A broad range of helping roles is considered. Studies selected for review include individual and group psychotherapy, crisis counseling, behavior modification, social and vocational rehabilitation programs, and academic-adjustment and mental-health-related services. Analogue or otherwise simulated therapeutic interactions and sensitivity or T-group experiences are omitted.

In keeping with traditional distinctions and definitions, individuals who have received postbaccalaureate, formal clinical training in professional programs of psychology, psychiatry, social work, and psychiatric nursing are considered to be professionals. Those who have not received this training are paraprofessionals.

The present review is divided into three sections. First, design criteria used to assess the methodology of comparative studies are described. Second, the characteristics and outcome of research studies are summarized, and the conclusions that present findings support are discussed. Third, experimental problems and limitations in current research are reviewed and suggestions for future work are offered.

The author wishes to thank Daryl "Kim" Hamblin for his help in the rating task described in this report.

Reprint requests should be sent to Joseph A. Durlak, Department of Psychology, Southern Illinois University, Carbondale, Illinois 62901.

Methodological Considerations

Luborsky, Singer, and Luborsky (1975) recently used 13 methodological criteria to evaluate the quality of research comparing alternative therapeutic approaches conducted by professionals. These criteria were modified for application to paraprofessional–professional comparative research. Two of the criteria used by Luborsky et al. regarding the need for experienced and mutually competent therapists were omitted, and two criteria, requiring multi-outcome and follow-up measures of client change, were added. The methodological criteria stressed such requirements as adequate sample size, equivalency of client groups, controls for expectancy effects and concurrent treatments, independent assessment of outcome, multi-outcome assessment of change, and at least one follow-up measure of client progress. For further details see Luborsky et al. (1975).

Forty-two studies comparing the effectiveness of professional and paraprofessional helpers were located. The experimental quality of each study was evaluated according to a 5-point index by applying the 13 design criteria. The grading system was not intended to be exact, but only to approximate the relative research sophistication of selected studies. As a partial check of the evaluation system, a second judge was asked to apply the criteria to five randomly selected studies. Interrater agreement was 92%.

Research Findings and Conclusions

Table 1 summarizes the characteristics, outcome, and experimental quality of the forty-two comparative studies. These studies represent a diversity of paraprofessional helpers, clinical settings, client populations, service programs, and target problems. Experienced psychologists, psychiatrists, and social workers typically constituted the professional therapist group. Only 10 studies exclusively used advanced clinical trainees as professional therapists (Elliott & Denney, 1975; Getz, Fujita, & Allen, 1975; Jensen, 1961; Kazdin, 1975; Levitz & Stunkard, 1974; Lindstrom, Balch, & Reese, 1976; Moleski & Tosi, 1976; Penick, Filion, Fox, & Stunkard, 1971; Ryan, Krall, & Hodges, 1976; Wolff, 1969).

The studies are grouped in Table 1 according to five major categories of helping services. The 19 studies in Group 1 involved individual and group psychotherapy or counseling primarily for moderately to severely disturbed adults; the four studies in Group 2 dealt with academic counseling for college students; the three studies in Group 3 involved crisis intervention for adults; and the 13 studies in Group 4 dealt with specific target problems of college students ($n = 5$), adults ($n = 6$), and children ($n = 2$), such as obesity, stuttering, insomnia, test and speech anxiety, and enuresis. Three studies fell into an *other* category (Group 5); these included Wolff's (1969) interpersonal training groups for normal college students, Schortinghuis and Frohman's (1974) cognitive tutoring program for handicapped preschool children, and Lamb and Clack's (1974) orientation program to a campus counseling center to increase use of this mental health resource among college students.

Thirty-five comparative studies have used multi-outcome measures of client change, and 27 studies have collected follow-up data on at least one measure. Representative outcome criteria have included performance on standardized psychological tests and various psychometric instruments (Lick & Heffler, 1977; Schortinghuis & Frohman, 1974), clients' self-reported change and satisfaction with services (Getz et al., 1975; Lamb & Clack, 1974), clinical ratings offered by independent judges (O'Brien, Hamm, Ray, Pierce, Luborsky, & Mintz, 1972), information from significant others (Miles, McLean, & Maurice, 1976; Wolff, 1969), academic or work performance (Mosher, Menn, & Matthews, 1975; Zunker & Brown, 1966), behavior ratings (Appleby, 1963; Ellsworth, 1968), analysis of therapist-offered empathy, warmth, and genuineness (Knickerbocker & McGee, 1973; Truax, 1967), performance in role-playing or in vivo situations (Fremouw & Harmatz, 1975; Moleski & Tosi, 1976), therapist improvement ratings (Karlsruher, 1976), supervisor evaluations (Covner, 1969; Magoon & Golann, 1966), criteria specific to treatment goals,

Table 1
Characteristics, Outcome, and Experimental Quality of Comparative Studies of Paraprofessional and Professional Helpers

Study	Experimental quality	Paraprofessional helpers	Client and helper sample size[a]	Results significantly favoring
Group 1: Individual or group psychotherapy or counseling				
Ellsworth (1968)	A	Psychiatric aides	327 psychiatric inpatients (?, ?)	Paraprofessionals
Jensen (1961)	B	Nurses and attendants	75 psychiatric inpatients[b] (?, 3)	Neither group
Karlsruher (1976)	B	College students	60 school children[b] (20, 6)	Neither group
Miles, McLean, & Maurice (1976)	B	Medical students	120 psychiatric inpatients (60, 27)	Neither group
O'Brien, Hamm, Ray, Pierce, Luborsky, & Mintz (1972)	B	Medical students	86 psychiatric outpatients (4, 12)	Neither group
Truax (1967)	B	Adult women	Over 300 vocational rehabilitation clients (4, 4)	Paraprofessionals
Truax & Lister (1970)	B	Adult women	168 vocational rehabilitation clients (4, 4)	Paraprofessionals
Weinman, Kleiner, Yu, & Tillson (1974)	B	Community volunteers	179 psychiatric outpatients (?, ?)	Neither group
Anker & Walsh (1961)	C	Occupational therapist	56 psychiatric inpatients (1, 1)	Paraprofessional
Appleby (1963)	C	Psychiatric aids	53 psychiatric inpatients[b] (?, ?)	Neither group
Colarelli & Siegel (1966)	C	Psychiatric aides	477 psychiatric inpatients (8, ?)	Neither group
Cole, Oetting, & Miskimins (1969)	C	Adult women	22 adolescent delinquents[b] (2, 2)	Neither group
Engelkes & Roberts (1970)	C	Adult counselors	1,502 vocational rehabilitation clients (142, 67)	Neither group
Mosher, Menn, & Matthews (1975)	C	Adult counselors	44 psychiatric inpatients (6, ?)	Paraprofessionals
Poser (1966)	C	College students	295 psychiatric inpatients[b] (11, 15)	Paraprofessionals
Sheldon (1964)	C	General physicians and nurses	83 psychiatric outpatients (?, ?)	Professionals better than physicians but equal to nurses
Mendel & Rapport (1963)	D	Psychiatric aides	166 psychiatric outpatients (?, ?)	Neither group
Covner (1969)	E	Community volunteers	Alcoholics[c] (?, ?)	Neither group
Magoon & Golann (1966)	E	Adult women	Psychiatric outpatients[c] (8, ?)	Neither group
Group 2: Academic counseling or advising for college students				
Zunker & Brown (1966)	A	College students	320 college students (8, 4)	Paraprofessionals
Brown & Myers (1975)	C	College students	303 college students (?, ?)	Neither group
Zultowski & Catron (1976)	C	College students	188 college students (10, ?)	Neither group
Murray (1972)	C	College students	166 college students (20, 9)	Neither group

Table 1 (*continued*)

Study	Experimental quality	Paraprofessional helpers	Client and helper sample size[a]	Results significantly favoring
Group 3: Crisis intervention for adults				
Knickerbocker & McGee (1973)	B	Community volunteers	92 adults and adolescents in crisis (65, 27)	Paraprofessionals
DeVol (1976)	E	Adult counselors	45 adults in crisis (4, 5)	Neither group
Getz, Fujita, & Allen (1975)	E	Community volunteers	104 adults in crisis (?, ?)	Neither group
Group 4: Interventions directed at specific target problems				
Kazdin (1975)	A	College students	54 unassertive adults and college students (?, ?)[d]	Neither group
Lick & Heffler (1977)	A	College student	40 adult insomniacs[b] (1, 1)	Neither group
Moleski & Tosi (1976)	A	Speech pathologist	20 adult stutterers[b] (1, 1)	Neither group
Elliott & Denney (1975)	B	College students	45 overweight college students[b] (3, 1)	Neither group
Levenberg & Wagner (1976)	B	Public health officer	54 adult smokers (1, 1)	Neither group
Levitz & Stunkard (1974)	B	Community volunteers	234 overweight adults[b] (8, 4)	Professionals
Lindstrom, Balch, & Reese (1976)	B	College students	68 overweight college students[b] (4, 1)	Neither group
Penick, Filion, Fox, & Stunkard (1971)	B	Adult volunteers	32 overweight adults (2, 1)	Neither group
Russell & Wise (1976)	B	College students	42 speech-anxious college students[b] (3, 3)	Neither group
Ryan, Krall, & Hodges (1976)	B	College students	72 test-anxious college students[b] (1, 2)	Neither group
Werry & Cohrssen (1965)	C	Parents	70 enuretic children[b] (22, 4)	Paraprofessionals
De Leon & Mandell (1966)	D	Parents	87 enuretic children[b] (56, 4)	Paraprofessionals
Fremouw & Harmatz (1975)	D	College students	30 speech-anxious college students[b] (11, 1)	Neither group
Group 5: Other interventions				
Lamb & Clack (1974)	B	College students	1,192 college students (?, 2)	Paraprofessionals
Schortinghuis & Frohman (1974)	C	Community volunteers	37 handicapped children (4, 3)	Paraprofessionals
Wolff (1969)	D	College students	88 college students[b] (4, 4)	Neither group

Note. A indicates that the design criteria were mainly satisfied; B, that one or two criteria were deficient; C, that three or four were deficient; D, that five were deficient; and E, that deficiencies were present in more than five criteria.
[a] Figures in parentheses are the number of paraprofessional and professional helpers, respectively; a ? indicates that the exact number of helpers was not specified.
[b] Includes no-treatment or attention-placebo control groups.
[c] Five therapists participated but a breakdown according to helper groups was not provided.
[d] Client sample size was not indicated.

such as weight loss or frequency of target behaviors (De Leon & Mandell, 1966; Levitz & Stunkard, 1974), and in hospital studies, such measures as length of hospital stay, ward transfers, discharge or readmission rates, and posthospital social and community adjustment (Jensen, 1961; Poser, 1966; Weinman, Kleiner, Yu, & Tillson, 1974).

Summaries of four studies involving psychiatric inpatients, children, adult insomniacs, and college students as clients are described below to provide the reader with some concrete examples of comparative research.

In a well-designed study, Zunker and Brown (1966) compared the effectiveness of four professional and eight student counselors who provided academic adjustment counseling to matched groups of 80 college freshmen. All counselors received 50 hours of identical training, used identical counseling materials, and followed an identical counseling sequence. Outcome criteria were assessments of study skills and study problems, academic grades, retention of information offered in counseling, and a counselor evaluation questionnaire. Student counselors were as effective as professionals on the first measure above but achieved significantly better results than did the professionals on each of the last four measures.

Lick and Heffler (1977) randomly assigned 40 adult insomniacs to four conditions: (a) therapist-administered progressive relaxation, (b) therapist-administered relaxation plus a taped relaxation program for client use at home, (c) placebo control, and (d) no-treatment control. A college undergraduate and a clinical psychologist saw clients in each treatment group. There were no outcome differences between the professional and paraprofessional therapists, as measured by client self-reported change and Minnesota Multiphasic Personality Inventory scores. The relaxation treatments did not differ significantly from each other, but both were significantly more effective than placebo and no-treatment controls.

In a child therapy study, 60 maladapting fifth and sixth graders were randomly assigned to a no-treatment control condition, treatment conducted by professional therapists (experienced school guidance counselors and psychology graduate students), or treatment by two groups of untrained college undergraduates who did or did not receive clinical supervision (Karlsruher, 1976). Treated children received 10 sessions of individual client-centered therapy. The California Test of Personality completed by the child, therapist improvement ratings, and a teacher-completed behavior checklist were used as outcome criteria. Children seen by supervised paraprofessionals improved the most, closely followed by children treated by professionals. Control children demonstrated more improvement than children seen by unsupervised paraprofessionals. Unexpectedly, ratings of therapist-offered empathy, warmth, and genuineness were negatively related to positive client change in the three treatment groups.

The progress of two matched samples of 60 psychiatric inpatients seen by medical student therapists, or psychiatric residents and fully trained psychiatrists was evaluated by Miles et al. (1976). Patients reported on changes in symptoms at the time of hospital discharge and again at 3-month follow-up. Independent ratings of improvement were also obtained from patients' significant others and family physicians at follow-up. From 40% to 88% of the patients demonstrated improvement on the outcome criteria at the different time points of evaluation, and there were no significant outcome differences between patients treated by the medical students, or psychiatrists and residents.

Overall, outcome results in comparative studies have favored paraprofessionals. In one study, professionals were more effective than one group of paraprofessional helpers but were equal in effectiveness to a second paraprofessional comparison group (Sheldon, 1964). In only one study were professionals significantly more effective than all paraprofessionals with whom they were compared (Levitz & Stunkard, 1974). In terms of measurable outcome, there were no significant differences among helpers in 28 investigations, but paraprofessionals were significantly more effective than professionals in 12 studies. The central finding from these comparative studies is that the clinical outcomes that paraprofes-

sionals achieve are equal to or significantly better than those obtained by professionals.

The provocative conclusion from these comparative investigations is that professionals do not possess demonstrably superior therapeutic skills, compared with paraprofessionals. Moreover, professional mental health education, training, and experience are not necessary prerequisites for an effective helping person.

In four studies (Anker & Walsh, 1961; De Leon & Mandell, 1966; Penick et al., 1971; Werry & Cohrssen, 1965) paraprofessionals have provided different treatment services from those provided by professionals, so it is not possible to assess therapist and treatment effects separately. Results could be due as much to the alternative treatment as to the use of paraprofessionals. Although these therapist–treatment confounds are a serious methodological drawback, the results of these four studies are heuristic. Paraprofessionals were found to be significantly more effective than professionals in three of these four studies (Anker & Walsh, 1961; De Leon & Mandell, 1966; Werry & Cohrssen, 1965) and such data challenge professionals to look more closely at the nature and efficacy of traditional mental health practices.

On closer scrutiny, results offer stronger support for paraprofessionals in some program areas than in others. The strongest support for paraprofessional effectiveness has come from the Group 4 studies, those directed at the modification of specific target problems. This group of studies contains the only one reporting results significantly favoring professionals (Levitz & Stunkard, 1974), but the other 12 studies showed no outcome differences among helper groups, and 10 of the 13 Group 4 studies were of A or B experimental quality (see Table 1 above).

The studies involving individual or group psychotherapy or counseling (Group 1) contained the largest number of studies favoring paraprofessionals over professionals (7), but only 3 of these 7 and 8 of the total group of 19 studies were relatively well controlled (of A or B quality). The lack of more rigorous research evaluation in Group 1 studies is unfortunate. These studies involved moderately to severely disturbed clients, and one can argue that such therapeutic programs are clinically more demanding than interventions attempting to modify discrete problem behaviors.

Academic counseling for college students, crisis intervention for adults, and the *other* interventions (Groups 2, 3, and 5, respectively) have not yet been carefully evaluated to offer any strong conclusions regarding paraprofessionals' clinical skills compared with those of professionals in these areas.

Current evidence offers reasonable, initial support for paraprofessionals' clinical effectiveness in comparative studies based on the following three considerations:

1. On the whole, the experimental quality of the studies in Table 1 approached or exceeded that observed in reviews of outcome research in other clinical areas (Bednar & Lawlis, 1971; Cash, 1973; Gurman, 1973; Kelley, Smits, Leventhal, & Rhodes, 1970; Luborsky, Chandler, Auerbach, Cohen, & Bachrach, 1971). Although some comparative reports were of little evaluative worth, others were relatively better controlled investigations on which at least tentative conclusions can be based.

2. Results are consistent regardless of the research sophistication of the study. Convergent evidence obtained by independent investigators using different design strategies and methods of evaluation lends strength to obtained findings.

3. Finally, several studies contained biases against paraprofessional treatment (Colarelli & Siegel, 1966; Fremouw & Harmatz, 1975; Jensen, 1961; Mendel & Rapport, 1963; Poser, 1966; Truax, 1967; Truax & Lister, 1970). Biases notwithstanding, positive findings of paraprofessionals' effectiveness suggest that a genuine clinical phenomenon is being observed.

Nevertheless, several reservations must be offered in evaluating the current status of comparative research. These issues are best discussed in reference to current experimental inadequacies and limitations and point the way toward future research.

Experimental Problems and Limitations

Design Criteria

The greatest deficiencies in design criteria were in the failure to assess the effects of additional, concurrent treatment (22 of 42 studies), to obtain multi-outcome (7 studies) or follow-up measures of client change (16 studies), or to rely on subjective estimates of treatment outcome (9 studies).

The criterion regarding the possibility of differential treatment expectations in the two helper-and-client groups was the most difficult to score. Although several authors noted biases against paraprofessional treatment (see above), a reverse bias was seldom noted. The initiation of a new treatment program, even if conducted by paraprofessionals, might create more positive expectations among staff and patients, in comparison with routine professional treatment. However, in most studies, it was difficult to determine whether this phenomenon was present.

The 13 design criteria adapted from Luborsky et al. (1975) are offered as a model to improve the methodology of future comparative research, with particular emphasis given to the most common deficiencies just mentioned. In addition to these design features, investigators should be aware of at least four other issues.

Patient Factors

The strongest evidence for paraprofessionals' effectiveness derives from studies of college students and adults with specific target problems (Group 4 studies) and moderately to severely disturbed non-middle-class adults (Group 1 studies). Work with adolescents is represented by a single study (Cole, Oetting, & Miskimins, 1969). Interventions with younger children concentrating on specific target behaviors have not been well controlled (De Leon & Mandell, 1966; Schortinghuis & Frohman, 1974; Werry & Cohrssen, 1965) and there has only been one study focusing on children's general adjustment difficulties (Karlsruher, 1976). No well-controlled comparative studies offering individual or group therapy have been conducted with clients, including college students, from the middle or upper socioeconomic classes. Since treatment effects might vary as a function of population characteristics, more comparative studies are needed dealing with the diverse problems of children, adolescents, college students, and adults not requiring psychiatric hospitalization or residential care.

Therapist Factors

Future work should examine the extent to which three therapist factors have contributed to current research findings—the use of unmatched therapist groups, the failure to study individual clinical performance, and the use of small therapist samples.

1. Paraprofessionals and professionals are frequently unmatched helper groups. In addition to formal clinical training and experience, paraprofessionals often differ from professionals in being younger, more often female, and from a lower socioeconomic class. The effects of these therapist variables in psychotherapy research are not clear (Meltzoff & Kornreich, 1970) but merit future study, especially in relation to therapy process dimensions and patient characteristics that may interact with outcome.

2. Individual clinical performance is seldom studied. It is possible that an emphasis on group performance obscures significant individual variability in helping effectiveness. By identifying certain workers who are particularly effective or ineffective within each helper group, the study of individual therapists' functioning might clarify the finding that on the average, paraprofessionals are able to do as well as professionals.

3. Finally, therapist samples in comparative studies are usually very small, yielding no indication of the representativeness of findings. Larger samples of paraprofessionals should be studied to provide information on the relative numbers and characteristics of personnel who can perform various clinical roles and functions adequately.

Selection Training and Supervision

Few studies have operationally defined their selection, training, and supervisory pro-

cedures, and fewer still have studied these program dimensions. Two studies comparing the presence versus absence of training and supervision obtained conflicting results. Karlsruher (1976) found that unsupervised and untrained college students were ineffective therapists for maladapting elementary school children, whereas untrained but supervised students and experienced professionals achieved equally successful results. In contrast, Lindstrom et al. (1976) reported that untrained and unsupervised college undergraduates were as effective as trained and supervised undergraduates and a professional therapist in helping college students lose weight.

Paraprofessionals have been selected on the basis of the results of psychological testing or a personal interview (Brown & Myers, 1975; Covner, 1969; Karlsruher, 1976; Russell & Wise, 1976), have been chosen because they were the working staff on clinical units or wards (Anker & Walsh, 1961; DeVol, 1976; Jensen, 1961; Miles et al., 1976; Moleski & Tosi, 1976), and have essentially been self-selected in the sense that available volunteers were accepted following apparently perfunctory screening procedures (Fremouw & Harmatz, 1975; Levitz & Stunkard, 1974; Murry, 1972; Poser, 1966). Most investigators, however, have not reported their selection process and criteria.

Some paraprofessionals have received no clinical training except a brief program orientation (Karlsruher, 1976; Lamb & Clack, 1974; Levenberg & Wagner, 1976; Poser, 1966; Weinman et al., 1974), some have received brief (up to 15 hours) training (Brown & Myers, 1975; Covner, 1969; Fremouw & Harmatz, 1975; Lindstrom et al., 1976; Levitz & Stunkard, 1974; Murry, 1972; Russell & Wise, 1976; Zultowski & Catron, 1976), and a few have participated in intensive programs approximating the training offered to professionals (Colarelli & Siegel, 1966; Magoon & Golann, 1966; Zunker & Brown, 1966). In most cases, only global descriptions of training programs have been presented. Both individual and group supervision have been used, but no one has detailed the supervisory process beyond clinical clichés.

Important parameters of selection, training, and supervision undoubtedly relate to program settings, treatment activities, goals, and client populations served. Judicious selection, training, and supervision might well account for paraprofessional effectiveness in comparative studies. Unfortunately, it is impossible to abstract the necessary details from comparative studies in order to analyze how paraprofessionals can be effectively selected, trained, and supervised. More systematic research should be devoted to these important program features.

The Process of Paraprofessional Intervention

Probably the most serious weakness in comparative research lies in the failure to examine the factors that account for paraprofessionals' effectiveness. Investigators have failed to relate specific intervention techniques to specific client changes. The nature of paraprofessional treatment is frequently left unclarified or is defined in global, undifferentiated terms. The nature of the paraprofessional's therapeutic influence is therefore undetermined.

Theories abound but evidence is meager about why nonprofessionals are effective helpers. For example, it is not known if paraprofessionals capitalize on powerful, nonspecific (placebo) therapeutic influences, use natural helping skills that perhaps reside in their interpersonal style, or adopt intervention techniques previously identified as effective in studies of professional therapists.

Paraprofessionals' higher interest and enthusiasm may make them as or more effective than professionals, but it seems unreasonable that these factors would work uniformly in all studies. Moreover, the perceived prestige and expertise often attributed to professionals by clients would probably minimize differential therapeutic effects accruing to paraprofessionals as a result of their enthusiasm and interest. Two studies found that paraprofessionals offered significantly higher levels of empathy, warmth, or genuineness to their clients than professionals did (Knickerbocker & McGee, 1973; Truax, 1967). These dimensions could account for paraprofessionals' effectiveness in these investigations and in

other programs as well. However, this issue is complex. Karlsruher (1976) reported that therapist-offered empathy, warmth, and genuineness were negatively related to positive changes occurring in child clients. Furthermore, there is controversy regarding the construct validity of ratings of empathy and other client-centered variables (Avery, D'Augelli, & Danish, 1976; Chinsky & Rappaport, 1970).

Paraprofessional effectiveness in some studies may be due to the development of carefully standardized and systematic treatment programs (Elliott & Denney, 1975; Kazdin, 1975; Levenberg & Wagner, 1976; Lick & Heffler, 1977; Lindstrom et al., 1976; Russell & Wise, 1976). In these programs, treatment has consisted of a programmed series of activities. Presumably, the more intervention procedures that can be clearly described and sequentially ordered in a helping program, the easier it will be for less trained personnel to administer them successfully. Paraprofessionals may feel more comfortable and hold higher expectations than professionals when using standardized clinical procedures, and these factors could contribute to paraprofessionals' clinical effectiveness. Therefore, paraprofessional clinical success may be closely related to professionals' abilities to define, order, and structure effective sequences of helping activities when training or supervising paraprofessionals. Although systematic treatment programs have been well controlled and have provided the strongest evidence of paraprofessionals' helping skills (Group 4 studies), paraprofessionals have exercised a wide latitude of clinical responsibility in most comparative studies and have not followed standardized and predetermined therapeutic procedures.

In summary, it is frustrating to admit that we do not know exactly why paraprofessionals with relatively little clinical experience and training can achieve results equal to or better than those obtained by professionals. Future research should attempt to define, isolate, and evaluate the primary treatment ingredients of paraprofessional helping programs. Research is needed in which the behavioral dimensions of nonprofessional intervention are described and assessed in relation to overall treatment effectiveness and specific client change.

The experimental limitations of comparative research must be discussed in relation to two important issues regarding the use of paraprofessionals within the mental health field. The first of these concerns the establishment of paraprofessional associate of arts programs and other new career training. The second deals with claims for the value of the indigenous therapist in reaching and helping various client populations.

Paraprofessional Mental Health Manpower

The intent of many undergraduate degree programs is to produce mental health generalists, that is, workers with a variety of basic helping skills who can eventually carry out many of the functions traditionally performed by professionals. Currently, over 4,000 graduates are being produced annually by approximately 170 paraprofessional college-level training programs (Young, True, & Packard, 1976). Present comparative evidence, however, does not support the ability of paraprofessionals to function effectively in full-time service positions.

Comparative research has studied workers' performance in very limited clinical roles with specific client groups. Paraprofessionals have primarily offered one form of treatment to homogeneous client populations in short-term experimental programs. In addition, research has neglected to assess paraprofessionals' and professionals' relative abilities to perform other primary clinical skills normally required in permanent service positions, such as intake interviewing, diagnostic assessment and evaluations, consultative services, and data collection for research purposes.

Therefore, comparative data supporting the value and wisdom of assimilating paraprofessionally trained workers as full-time staff in the human services are not currently available and can be obtained only in more comprehensive clinical studies. Systematic information is needed about the long-term effectiveness of paraprofessionals, employing a variety of intervention techniques with diverse client populations. We also need information on the relative abilities of paraprofes-

sionals and professionals to perform other primary clinical skills, ranging from intake interviewing to consultation. Pilot investigations examining workers' relative skills in these areas have only recently begun (Gingerich, Feldman, & Wodarski, 1976; Sloop & Quarrick, 1974).

Indigenous Therapists

Comparative research has offered only partial support for the value of indigenous therapists, that is, workers similar to clients in background, life-style, and general personal and demographic characteristics. The only controlled evidence for the effectiveness of indigenous therapists has involved college students (Brown & Myers, 1975; Elliott & Denney, 1975; Fremouw & Harmatz, 1975; Karlsruher, 1976; Lamb & Clack, 1974; Lindstrom et al., 1976; Murry, 1972; Russell & Wise, 1976; Ryan et al., 1976; Wolff, 1969; Zultowski & Catron, 1976; Zunker & Brown, 1966).

However, the indigenous therapist has often been cited as the treatment agent of choice for low-income and minority group clients who have not received a fair distribution of services from professionals in the past. It is believed that indigenous therapists can establish rapport and identification with previously underserved populations, which makes them more effective than professionals working with the same groups. Unfortunately, there are no experimental data to support this assertion. In fact, data from two analogue and attitude studies indicate that Mexican-American clients perceive white professionals to be as trustworthy, understanding, and helpful as indigenous therapists (Acosta & Sheehan, 1976; Andrade & Burstein, 1973). The comparative effectiveness of indigenous and professional helpers working with noncollege populations awaits empirical documentation.

Summary

Findings from 42 studies comparing the helping effectiveness of paraprofessionals and professionals are consistent and provocative. The clinical outcomes paraprofessionals achieve are equal to or significantly better than those obtained by professionals. These data suggest that professionals do not necessarily possess demonstrably superior clinical skills, in terms of measurable outcome, when compared with paraprofessionals. Moreover, professional mental health education, training, and experience are not necessary prerequisites for an effective helping person. The strongest support for paraprofessionals has come from programs directed at the modification of adults' and college students' specific target problems and, to a lesser extent, from group and individual therapy programs for non-middle-class adults.

Although the above findings and conclusions must be offered tentatively due to deficiencies and limitations in the methodology of many studies, the consistency of positive findings supports the potential value of paraprofessional helpers. A set of 13 design criteria adapted from Luborsky et al. (1975) is offered as a guide to improve the experimental rigor of future research. Future investigators should pay particular attention to a number of unresolved issues:

1. What are the primary treatment ingredients in paraprofessional helping programs? What specific behavioral dimensions of paraprofessional intervention are associated with treatment effectiveness?
2. How variable is individual clinical effectiveness within paraprofessional helper groups? Is a minority or a majority of helpers responsible for the positive findings that have been obtained? Do individual differences in clinical success relate to client, treatment, or therapist characteristics?
3. What is the breadth of paraprofessionals' clinical competence? Are paraprofessionals effective on a long-term basis when working with diverse client populations and providing more than one form of treatment or service? Can paraprofessionals adequately perform those therapy-related skills that would be required of them in full-time service positions such as clinical interviewing and diagnostic assessments?
4. Are indigenous therapists the treatment agent of choice for any client populations? If so, what factors account for this finding? For example, in comparison with professionals,

is the supposition correct that indigenous helpers are more accepted by and more empathic toward clients similar to themselves? Furthermore, are these variables associated with therapeutic outcome?

5. Finally, what are the most effective means of selecting, training, and supervising paraprofessionals? Generally, attempts in these directions have been nonspecific in comparative research, due to the limited number of therapists studied. However, if future work continues to confirm paraprofessionals' treatment effectiveness in comparative situations, then the most judicious use of paraprofessionals will, in large part, depend on determining the most successful techniques for recruiting, training, and supervising these new workers.

This review has used a broad definition of helping that encompasses a wide range of settings, helpers, clients, treatments, and criterion variables. This generic assessment of paraprofessional functioning seems appropriate given the current lack of experimentally rigorous process and outcome research. A more limited review of specific treatment situations would not have afforded any greater scrutiny or insight into paraprofessionals' abilities. However, as future investigators achieve stricter control of therapist, treatment, and client characteristics, paraprofessionals' functioning under different circumstances can be assessed more accurately and reliably in order to challenge or qualify the general conclusions offered here.

Current findings are not offered as a polemic against professional treatment but as a stimulus to investigate the processes that facilitate behavior change. Data indicate that paraprofessionals can make an important contribution as helping agents, but the factors accounting for this phenomenon are not understood. We are presently recruiting and employing thousands of paraprofessionals in the human services without an adequate understanding of why such personnel are effective helpers. It would be a mistake to continue using paraprofessionals without more closely examining their skills, deficiencies, and limitations.

References

Acosta, F. X., & Sheehan, J. G. Preferences toward Mexican American and Anglo American psychotherapists. *Journal of Consulting and Clinical Psychology,* 1976, *44,* 272–279.

Andrade, S. J., & Burstein, A. G. Social congruence and empathy in paraprofessional and professional mental health workers. *Community Mental Health Journal,* 1973, *9,* 388–397.

Anker, J. M., & Walsh, R. P. Group psychotherapy, a special activity program, and group structure in the treatment of chronic schizophrenics. *Journal of Consulting Psychology,* 1961, *25,* 476–481.

Appleby, L. Evaluation of treatment methods for chronic schizophrenia. *Archives of General Psychiatry,* 1963, *8,* 8–21.

Avery, A. W., D'Augelli, A. R., & Danish, S. J. An empirical investigation of the construct validity of "empathic understanding" ratings. *Counselor Education and Supervision,* 1976, *15,* 177–183.

Bednar, R. L., & Lawlis, G. F. Empirical research in group psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change.* New York: Wiley, 1971.

Berkowitz, B. P., & Graziano, A. M. Training parents as behavior therapists: A review. *Behaviour Research and Therapy,* 1972, *10,* 297–317.

Brown, C. R., & Myers, R. Student vs. faculty curriculum advising. *Journal of College Student Personnel,* 1975, *16,* 226–231.

Cash, T. F. Methodological problems and progress in schizophrenia research: A survey. *Journal of Consulting and Clinical Psychology,* 1973, *40,* 278–286.

Chinsky, J. M., & Rappaport, J. Brief critique of the meaning and reliability of "accurate empathy" ratings. *Psychological Bulletin,* 1970, *73,* 379–382.

Colarelli, N. J., & Siegel, S. M. *Ward H: An adventure in innovation.* Princeton, N.J.: Van Nostrand, 1966.

Cole, C. W., Oetting, E. R., & Miskimins, R. W. Self-concept therapy for adolescent females. *Journal of Abnormal Psychology,* 1969, *74,* 642–645.

Covner, B. J. Screening volunteer alcoholism counselors. *Quarterly Journal of Studies on Alcohol,* 1969, *30,* 420–425.

De Leon, G., & Mandell, W. A comparison of conditioning and psychotherapy in the treatment of functional enuresis. *Journal of Clinical Psychology,* 1966, *22,* 326–330.

DeVol, T. I. Does level of professional training make a difference in crisis intervention counseling? *Journal of Community Health,* 1976, *2,* 31–35.

Elliott, C. H., & Denney, D. R. Weight control through covert sensitization and false feedback. *Journal of Consulting and Clinical Psychology,* 1975, *43,* 842–850.

Ellsworth, R. B. *Nonprofessionals in psychiatric rehabilitation.* New York: Appleton-Century-Crofts, 1968.

Engelkes, J. R., & Roberts, R. R. Rehabilitation counselors' level of training and job performance. *Journal of Counseling Psychology*, 1970, *17*, 522–526.

Fremouw, W. J., & Harmatz, M. G. A helper model for behavioral treatment of speech anxiety. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 652–660.

Getz, W. L., Fujita, B. N., & Allen, D. The use of paraprofessionals in crisis intervention: Evaluation of an innovative program. *American Journal of Community Psychology*, 1975, *3*, 135–144.

Gingerich, W. J., Feldman, R. A., & Wodarski, J. S. Accuracy in assessment: Does training help? *Social Work*, 1976, *21*, 40–48.

Gruver, G. G. College students as therapeutic agents. *Psychological Bulletin*, 1971, *76*, 111–127.

Gurman, A. S. The effects and effectiveness of marital therapy: A review of outcome research. *Family Process*, 1973, *12*, 145–170.

Jensen, M. B. Consultation vs therapy in the psychological treatment of NP hospital patients. *Journal of Clinical Psychology*, 1961, *17*, 265–268.

Johnson, C. A., & Katz, R. C. Using parents as change agents for their children: A review. *Journal of Child Psychology and Psychiatry*, 1973, *14*, 181–200.

Karlsruher, A. E. The nonprofessional as a psychotherapeutic agent: A review of the empirical evidence pertaining to his effectiveness. *American Journal of Community Psychology*, 1974, *2*, 61–77.

Karlsruher, A. E. The influence of supervision and facilitative conditions on the psychotherapeutic effectiveness of nonprofessional and professional therapists. *American Journal of Community Psychology*, 1976, *4*, 145–154.

Kazdin, A. E. Covert modeling, imagery assessment, and assertive behavior. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 716–724.

Kelley, J., Smits, S. J., Leventhal, R., & Rhodes, R. Critique of the designs of process and outcome research. *Journal of Counseling Psychology*, 1970, *17*, 337–341.

Knickerbocker, D. A., & McGee, R. K. Clinical effectiveness of nonprofessional and professional telephone workers in a crisis intervention center. In D. Lester & G. Brockopp (Eds.), *Telephone therapy and crisis intervention*. Springfield, Ill.: Charles C Thomas, 1973.

Lamb, D. H., & Clack, R. J. Professional versus paraprofessional approaches to orientation and subsequent counseling contacts. *Journal of Counseling Psychology*, 1974, *21*, 61–65.

Levenberg, S. B., & Wagner, M. K. Smoking cessation: Long-term irrelevance of mode of treatment. *Journal of Behavior Therapy and Experimental Psychiatry*, 1976, *7*, 93–95.

Levitz, L. S., & Stunkard, A. J. A therapeutic coalition for obesity: Behavior modification and patient self-help. *American Journal of Psychiatry*, 1974, *131*, 423–427.

Lick, J. R., & Heffler, D. Relaxation training and attention placebo in the treatment of severe insomnia. *Journal of Consulting and Clinical Psychology*, 1977, *45*, 153–161.

Lindstrom, L. L., Balch, P., & Reese, S. In person versus telephone treatment for obesity. *Journal of Behavior Therapy and Experimental Psychiatry*, 1976, *7*, 367–369.

Luborsky, L., Chandler, M., Auerbach, A. H., Cohen, J., & Bachrach, H. M. Factors influencing the outcome of psychotherapy: A review of quantitative research. *Psychological Bulletin*, 1971, *75*, 145–185.

Luborsky, L., Singer, B., & Luborsky, L. Comparative studies of psychotherapies. *Archives of General Psychiatry*, 1975, *32*, 995–1008.

Magoon, T. M., & Golann, S. E. Nontraditionally trained women as mental health counselors/psychotherapists. *Personnel and Guidance Journal*, 1966, *44*, 788–793.

Meltzoff, J., & Kornreich, M. *Research in psychotherapy*. New York: Atherton Press, 1970.

Mendel, W. M., & Rapport, S. Outpatient treatment for chronic schizophrenic patients: Therapeutic consequences of an existential view. *Archives of General Psychiatry*, 1963, *8*, 190–196.

Miles, J. E., McLean, P. D., & Maurice, W. L. The medical student therapist: Treatment outcome. *Canadian Psychiatric Association Journal*, 1976, *21*, 467–472.

Moleski, R., & Tosi, D. J. Comparative psychotherapy: Rational-emotive therapy versus systematic desensitization in the treatment of stuttering. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 309–311.

Mosher, L. R., Menn, A., & Matthews, S. M. Soteria: Evaluation of a home-based treatment for schizophrenia. *American Journal of Orthopsychiatry*, 1975, *45*, 455–467.

Murry, J. P. The comparative effectiveness of student-to-student and faculty advising programs. *Journal of College Student Personnel*, 1972, *13*, 562–566.

O'Brien, C. P., Hamm, K. B., Ray, B. A., Pierce, J. F., Luborsky, L., & Mintz, J. Group vs individual psychotherapy with schizophrenics: A controlled outcome study. *Archives of General Psychiatry*, 1972, *27*, 474–478.

O'Dell, S. Training parents in behavior modification: A review. *Psychological Bulletin*, 1974, *81*, 418–433.

Penick, S. B., Filion, R., Fox, S., & Stunkard, A. Behavior modification in the treatment of obesity. *Psychosomatic Medicine*, 1971, *33*, 49–55.

Poser, E. G. The effects of therapists' training on group therapeutic outcome. *Journal of Consulting Psychology*, 1966, *30*, 283–289.

Russell, R. K., & Wise, F. Treatment of speech anxiety by cue-controlled relaxation and desensitization with professional and paraprofessional counselors. *Journal of Counseling Psychology*, 1976, *23*, 583–586.

Ryan, V. L., Krall, C. A., & Hodges, W. F. Self-concept change in behavior modification. *Journal of Consulting and Clinical Psychology,* 1976, *44,* 638–645.

Schortinghuis, N. E., & Frohman, A. A comparison of paraprofessional and professional success with preschool children. *Journal of Learning Disabilities,* 1974, *7,* 245–247.

Sheldon, A. An evaluation of psychiatric after-care. *British Journal of Psychiatry,* 1964, *110,* 662–667.

Siegel, J. M. Mental health volunteers as change agents. *American Journal of Community Psychology,* 1973, *1,* 138–158.

Sloop, E. W., & Quarrick, E. Technician performance: Reliability and validity. *Professional Psychology,* 1974, *5,* 216–218.

Truax, C. B. The use of supportive personnel in rehabilitation counseling: Process and outcome. In G. B. Leslie (Ed.), *Supportive personnel in rehabilitation centers: Current practices and future needs.* Washington, D.C.: Association of Rehabilitation Centers, 1967.

Truax, C. B., & Lister, J. L. Effectiveness of counselors and counselor aides. *Journal of Counseling Psychology,* 1970, *17,* 331–334.

Weinman, B., Kleiner, R., Yu, J. H., & Tillson, V. A. Social treatment of the chronic psychotic patient in the community. *Journal of Community Psychology,* 1974, *2,* 358–365.

Werry, J. S., & Cohrssen, J. Enuresis—An etiologic and therapeutic study. *Journal of Pediatrics,* 1965, *67,* 423–431.

Wolff, T. Undergraduates as campus mental health workers. *Personnel and Guidance Journal,* 1969, *48,* 294–304.

Young, C. E., True, J. E., & Packard, M. E. A national study of associate degree mental health and human services workers. *Journal of Community Psychology,* 1976, *4,* 89–95.

Zultowski, W. H., & Catron, D. W. Students as curriculum advisers: Reinterpreted. *Journal of College Student Personnel,* 1976, *17,* 199–204.

Zunker, V. G., & Brown, W. F. Comparative effectiveness of student and professional counselors. *Personnel and Guidance Journal,* 1966, *44,* 738–743.

Received September 26, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 93–118

Learned Helplessness in Humans: A Review and Attribution-Theory Model

Ivan W. Miller III
Brown University and Butler Hospital
Providence, Rhode Island

William H. Norman
Section of Psychiatry and Human Behavior
Brown University and Butler Hospital
Providence, Rhode Island

Seligman's theory of learned helplessness and the current status of the research literature are reviewed, with a focus on five issues of the learned helplessness phenomenon: (a) nature, (b) etiology, (c) generalization, (d) individual differences, and (e) alleviation. Seligman's theory is seen as inadequate to account for present data in several areas, notably etiology and generalization. A revised model of learned helplessness in humans is presented that suggests that the individual's attributions of noncontingent failure experiences predict the degree and parameters of learned helplessness.

Seligman and Maier (1967) and Overmier and Seligman (1967) used the term *learned helplessness* to describe an interference with escape–avoidance behaviors produced in dogs by prior inescapable shock. Since these early studies, research bearing on learned helplessness has proliferated. Early work investigating the parameters of the phenomenon used dogs as subjects (Overmier, 1968; Seligman & Groves, 1970; Seligman, Maier, & Geer, 1968), and more recent studies have reported the occurrence of learned helplessness in cats (Thomas & Balter, in press), fish (Padilla, Padilla, Ketterer, & Giacolone, 1970), and rats (Braud, Wepman, & Russo, 1969; Maier, Seligman, & Soloman, 1969; Maier & Testa, 1975).

The dominant researcher and theorist in this area has been Martin E. P. Seligman, who has written extensively on the nature, etiology, and significance of the learned helplessness phenomenon (Seligman, 1973, 1974, 1975). Seligman's model[1] has broadened the scope of learned helplessness from animal behavior to a wide variety of human behaviors, including reactive depression, stomach ulcers, voodoo deaths, and child development (Seligman, 1975). Since the publication of the early animal work and the major exposition of Seligman's theory (Seligman, 1975), a substantial body of research has investigated the parameters of the learned helplessness phenomenon with human subjects. Recent reviews (Maier & Seligman, 1976) and critiques (Levis, 1976) of learned helplessness research and theory have focused primarily on animal research. In view of the considerable interest in this area of research and the

An earlier version of this manuscript was submitted by Ivan W. Miller III in partial fulfillment of the requirements for the doctoral degree at the University of Maine, Orono. Committee members Joel Gold, Richard Ryckman, and Gordon Kulberg made helpful comments on that preliminary draft.

The authors would like to thank R. Eric Nelson for his thoughtful comments on the final draft of this article.

Requests for reprints should be sent to Ivan W. Miller III, Butler Hospital, 345 Blackstone Boulevard, Providence, Rhode Island 02906.

[1] Seligman has recently presented a revised model of learned helplessness (Seligman, Note 1; Abramson, Seligman, & Teasdale, 1978). Since this revised model was presented during the preparation of the final draft of the present article, all references in the text to "Seligman's model" refer to his pre-1977 position. It should also be noted that the basic statements of Seligman's revised model are remarkably similar to those of the model presented in the second half of the present article. The present authors and Seligman (Note 2) regard this theoretical overlap to be the result of parallel, independent contributions; as such, these represent a situation that is unusual in psychology. The present article makes no attempt to review, critique, or integrate Seligman's revised theory with the model proposed here.

possible implications for human behavior, it seems appropriate at this time to review the learned helplessness research and theory with an explicit focus on studies using human subjects. The purpose of this article, then, is threefold: (a) to review the research literature concerning the phenomenon of learned helplessness in humans, (b) to assess the validity of Seligman's theory in view of recent research, and (c) to propose a new model of learned helplessness in humans.

In order to facilitate comparison of theoretical predictions and research findings, this article focuses on five salient issues concerning learned helplessness. (a) What is the nature of learned helplessness? (b) What are the necessary and sufficient conditions for development of learned helplessness? (c) Is learned helplessness a generalized human phenomenon, or is it confined to specific laboratory conditions? (d) What are the individual differences factors that affect learned helplessness? (e) What are the necessary and sufficient conditions for alleviation of learned helplessness? This article reviews the relevant research and theory for each of the above issues.

Before moving to an evaluation of the research findings, however, it is necessary to discuss the basic experimental paradigm used in learned helplessness research. In the typical study, subjects receive a *training phase* followed by a *test phase*. In the training phase, subjects are exposed to a *training task,* in which they receive (a) contingent (response-dependent) reinforcement, (b) noncontingent (response-independent) reinforcement, or (c) no treatment (control). After this training phase, the performance of the three groups is compared on a *test task,* in which reinforcement is contingent for all groups. Learned helplessness occurs when the subjects receiving noncontingent reinforcement in the training phase show deficits in the test phase relative to the contingent and control groups. Thus, learned helplessness refers to behavioral deficits produced by exposure to noncontingent outcomes.

The Nature of Learned Helplessness

Seligman (1975) suggested that learned helplessness consists of three interrelated areas of disturbance: (a) motivational, (b) cognitive, and (c) emotional. More specifically, Seligman hypothesized that learned helplessness "(1) reduces the motivation to control the outcome; (2) interferes with learning that responding controls the outcome; (3) produces fear for as long as the subject is uncertain of the uncontrollability of the outcome, and then produces depression" (Seligman, 1975, p. 56). In an attempt to explore the nature of learned helplessness and the validity of Seligman's hypotheses, this section of the article examines the changes in test-task performance reported in the learned helplessness research with human subjects. Thus, this section is concerned with the nature of learned helplessness as operationalized by performance in the test phase, and not with the training-phase conditions necessary to produce learned helplessness.

Cognitive and Motivational Deficits

We are aware of 23 studies that have attempted to demonstrate the occurrence of changes in performance due to a training phase in which the subjects' responses were independent of environmental outcomes. Two generalizations concerning the measurement of the changes in performance are immediately evident from a review of these studies. First, in an attempt to operationalize Seligman's proposed cognitive and motivational deficits, most studies have used a test task requiring development and use of cognitive problem-solving strategies. A smaller number of studies have investigated the emotional aspects of learned helplessness. Second, although a majority of these studies have reported performance deficits in the test phase, several studies (Roth & Bootzin, 1974; Roth & Kubal, 1975; Thornton & Jacobs, 1972) have reported increases in performance following a learned helplessness training phase. Since the relevant issues concerning this *facilitation effect* seem to lie in the training phase of learned helplessness, these studies are covered in the next section during the discussion of the necessary and sufficient conditions for the development of learned helplessness. The present section examines those studies that have reported significant performance deficits.

Five studies have attempted to replicate the type of task used in the animal research by utilizing an escape–avoidance task in the test phase and have found deficits with human subjects (Hiroto, 1974; Hiroto & Seligman, 1975; Klein & Seligman, 1976; Krantz, Glass, & Snyder, 1974; Thornton & Jacobs, 1971). Seven other studies have used an anagram-solution task (Benson & Kennelly, 1976; Gatchel, Paulus, & Maples, 1975; Gatchel & Proctor, 1976; Hiroto & Seligman, 1975; Klein, Fencil-Morse, & Seligman, 1976; Miller & Seligman, 1975; Miller & Gold, Note 3). In this task, subjects are given a series of anagrams with the same solution order. In both the escape–avoidance and anagram tasks, three dependent measures have generally been employed: (a) number of trials to escape (anagram solution) criterion, (b) number of failures to escape (solve anagram), and (c) mean escape (anagram solution) latency. The measure of number of trials to criterion was hypothesized to operationalize the cognitive deficit, and the latter two measures were hypothesized to operationalize the motivational deficit. However, as was noted by Miller and Seligman (1975), because solution criteria were defined in terms of response speed, separation of motivational and cognitive components was not possible.

It is also important to note that the absolute magnitude of the deficits obtained in learned helplessness studies using escape–avoidance or anagram test tasks was relatively small. In the Hiroto and Seligman (1975) escape–avoidance task, for example, the mean response latency for the noncontingent group was about 8 sec, whereas the mean latency for the control group was about 7.3 sec. Although statistically significant, the small absolute difference raises questions about the importance of these results.

Two studies (Klein & Seligman, 1976; Miller & Seligman, 1975) designed to isolate Seligman's hypothesized cognitive deficit have employed Rotter, Liverant, and Crowne's (1961) measure of expectancy change following success or failure. In this task, subjects are given a "chance" or "skill" task, in which reinforcement is covertly manipulated by the experimenter. The dependent measure is the amount of expectancy change after trials of success or failure. Several studies (Phares, 1957; Rotter et al., 1961) have reported that reinforcement on previous trials had a greater effect on expectancies for future success when the subjects perceived reinforcement as response contingent. Unfortunately, the usefulness of this measure for learned helplessness depends on the previous research suggesting that expectancy change in this task is a function of the expectancy of response–outcome contingency. Other research investigating expectancy changes in this task (McMahan, 1973; Weiner, Nierenberg, & Goldstein, 1976) has suggested that expectancy changes following success and failure are not due to perceived response–outcome contingency but are due to the perceived stability of the causal attributions of performance. Hence, the use of the expectancy-change measure devised by Rotter et al. to operationalize an expectancy of response–outcome contingency remains highly questionable. This issue, as well as the construct of attributions, is examined in detail in a later section of this article.

Although other studies have used a variety of cognitive problem-solving tasks, including intelligence tests (Thornton & Jacobs, 1972), block design (Dweck & Reppucci, 1973), digit–letter substitution (Dweck & Bush, 1976), discrimination learning (Eisenberger, Park, & Frank, 1976), and specially designed concept-formation problems (Roth & Bootzin, 1974; Roth & Kubal, 1975), no study has adequately separated motivational and cognitive components of learned helplessness.

In summary, the literature pertaining to the performance changes produced by exposure to learned helplessness training conditions supports the conceptualization of learned helplessness as a performance deficit in cognitive problem-solving tasks. This deficit, though statistically significant, is relatively small in an absolute sense. The available evidence does not allow a distinction between cognitive and motivational explanations for this deficit. Thus, the performance deficits that are defined as learned helplessness may have a cognitive or motivational basis, or they may result from the impairment of both processes.

Emotional Deficits

Six studies have investigated the emotional aspects of learned helplessness (Gatchel et al., 1975; Gatchel & Proctor, 1976; Griffith, 1977; Krantz et al., 1974; Miller & Seligman, 1975; Roth & Kubal, 1975). These studies generally support Seligman's hypothesis that learned helplessness involves feelings of anxiety and depression. Miller and Seligman (1975) and Gatchel et al. (1975) administered the Multiple Affect Adjective Check List (Zuckerman, Lubin, & Robins, 1965) before and after exposure to contingent and noncontingent reinforcement. Both studies reported significant increases in feelings of depression, anxiety, and hostility following noncontingent reinforcement. Griffith (1977) administered the Paired Anxiety and Depression Scale (Mould, 1975) before and after exposure to noncontingent conditions and reported significant increases in depression following noncontingent failure and significant increases in anxiety following noncontingent success. Similarly, Roth and Kubal (1975) and Krantz et al. (1974) administered questionnaires following the training phase and reported increases in feelings of helplessness, incompetence, stress, frustration, hostility, depression, anger, and fatigue. Two studies have investigated the emotional aspects of learned helplessness through physiological measures (Gatchel & Proctor, 1976; Krantz et al., 1974). Both studies reported that subjects exposed to learned helplessness training conditions showed lower levels of electrodermal activity, which is thought to be evidence of a lowered motivational state (Malmo, 1965), which has also been suggested to be associated with clinical depression (McCarron, 1973).

Thus, although the self-report and physiological data support Seligman's predictions of increased depression and anxiety following exposure to a learned helplessness induction, the self-report studies also suggest that exposure to these conditions results in increased hostility, a phenomenon neither predicted nor explained by Seligman's theory.

In summary, the available research generally supports Seligman's hypotheses concerning the nature of learned helplessness. Performance deficits appear attributable to cognitive and motivational deficits, although assessment of the relative contributions of these two processes has not been demonstrated. Similarly, Seligman's hypothesis concerning the emotional aspects of learned helplessness appears to be supported, although the observed increases in hostility were not predicted.

Two basic issues remain concerning the nature of learned helplessness. The first concerns the limited number of types of tasks that have been used in learned helplessness test phases. With a few exceptions, these tasks have been cognitive problem-solving tasks. Few other types of tasks have been used to investigate the effects of learned helplessness in other types of situations. The second issue concerns the degree of impairment found in human subjects. All of the helplessness effects reported with humans have been relatively small, with no reports of any behavior as disabling as was found in the earlier research with infrahumans. There are certainly ethical and legal reasons for not attempting to instill these deficits in humans, but the consistent findings of small difference between groups raises questions concerning the relative significance of learned helplessness in humans. Animal research may provide a reasonable facsimile, or it may not. It is certainly possible that the superior cognitive strategies available to humans prevent the extreme effects of learned helplessness. This question of the relevance of animal research for a theory of learned helplessness is discussed in a later section.

Etiology

Seligman (1973, 1974, 1975) has postulated that the major causal factor in the development of learned helplessness is the organism's belief or expectancy that its responses will not influence the future probability of environmental outcomes (expectancy of response–outcome independence). He has further proposed that information about this contingency is a property of the environment, not of the organism, but that this information is transformed into an expectancy of the response–outcome contin-

gency, and it is this internal expectancy that directs behavior. In the learned helplessness paradigm, it is only when the organism forms the expectancy that its response will not be effective that learned helplessness is predicted to occur. Seligman has also briefly mentioned three variables that may limit the acquisition of the expectancy of response independence and of learned helplessness: (a) previous response–outcome expectancies, (b) discrimination between situations, and (c) the relative importance of the situation. He has not discussed these variables in any detail, however, and has not integrated them into his general theoretical framework.

Before moving on to an examination of the research bearing on Seligman's theory, the purpose of this section should be clarified. In terms of the general experimental paradigm discussed in the introduction, this section is concerned with the training phase, that is, those experimental manipulations and conditions that have been used to operationalize Seligman's construct of response–outcome independence, and with the success of these manipulations in producing deficits in test-task performance.

Turning now to an examination of the research, all of the animal studies and over half of the studies using human subjects have used an uncontrollable aversive stimulus, either noise or shock, in the training phase of the learned helplessness experiment (Fosco & Geer, 1971; Gatchel et al., 1975; Gatchel & Proctor, 1976; Geer, Davison, & Gatchel, 1970; Glass & Singer, 1972; Hiroto, 1974; Hiroto & Seligman, 1975; Klein & Seligman, 1976; Krantz et al., 1974; Miller & Seligman, 1975, 1976; Sherrod & Downs, 1974; Thornton & Jacobs, 1971, 1972). In these studies, subjects were exposed to an instrumental escape–avoidance task. The contingent group could actually avoid or escape the aversive stimulus, but the noncontingent group's responses did not influence the presentation of the aversive stimulus. Although several of these studies have not used appropriate control groups (Wortman & Brehm, 1975), all studies using an uncontrollable aversive stimulus, except that by Thornton and Jacobs (1972), have reported that the noncontingent group showed significant performance deficits in the test phase. These studies offer support for Seligman's hypothesis by demonstrating that close replication of the experimental procedures used to induce learned helplessness in infrahumans will also lead to learned helplessness in humans.

However, these studies have provided only a partial test of the ramifications of Seligman's theory. Seligman (1975) suggested that learned helplessness will result not only from noncontingent aversive stimulation but from any noncontingent environmental outcomes, including positive reinforcement. One group of studies (Benson & Kennelly, 1976; Hiroto & Seligman, 1975; Roth & Bootzin, 1974; Roth & Kubal, 1975; Miller & Gold, Note 3) has attempted to test this more general hypothesis by inducing learned helplessness by exposure to noncontingent positive reinforcement. All of these studies have used the procedure first reported by Roth and Bootzin (1974), who exposed subjects to contingent or random (noncontingent) reinforcement concerning their performance on a series of concept-formation discrimination tasks, as were described by Levine (1971). Roth and Bootzin argued that the condition of random reinforcement corresponded to Seligman's theoretical condition of independence of responses and outcomes but was not confounded by the effects of exposure to aversive stimulation. The results of studies using this type of training condition have been mixed. Hiroto and Seligman (1975) and Benson and Kennelly (1976) reported that random-reinforcement groups showed deficits on an anagram test task, relative to contingent-reinforcement and control groups. Conversely, Roth and Bootzin (1974) reported a facilitation of performance following exposure to this type of training task. Roth and Kubal (1975) varied the number of tasks in which the subjects received random reinforcement. They reported that subjects exposed to only one learned helplessness training task showed facilitation effects, whereas subjects exposed to random reinforcement in two different training tasks showed deficits on the same test task. These results, however, have not been replicated (Kilpatrick-Tabak & Roth, Note 4). It should be noted that the test tasks used by

Roth and Bootzin (1974) and Roth and Kubal (1975) differed from those used by Hiroto and Seligman (1975) and Benson and Kennelly (1976) on several dimensions, notably on the similarity of the training and test phases. (These differences are discussed more thoroughly in a later section.)

In general, it is apparent that the results from studies using a random-reinforcement procedure have been far from definitive. Furthermore, the use of random reinforcement as noncontingent positive reinforcement is questionable. Benson and Kennelly (1976) cogently argued that since nonreward in the context of reward produces frustration, an aversive event (Amsel, 1972), a random schedule of rewards means that reward and frustration may also occur in a random and uncontrollable manner. A random-reinforcement procedure can thus be seen as confounding noncontingent positive reinforcement with noncontingent aversive stimulation.

Not only does a random-reinforcement procedure confound the type of noncontingent outcome, it also confounds the absolute amounts of positive and negative outcomes that occur. Studies using a random-reinforcement group have typically not yoked the noncontingent (random) group's reinforcement to the contingent group's level, as was done in the early animal studies, but instead have fixed the random group's reinforcement rate at 50%. Thus, the obtained differences between groups may be due to differences in the amount and pattern of reinforcement received.

Two studies (Benson & Kennelly, 1976; Miller & Gold, Note 3) have controlled for the amount and type of reinforcement. Miller and Gold used three groups in the standard Levine (1971) discrimination-problem training task: (a) contingent (80% correct), (b) yoked noncontingent (80% correct), and (c) random (50% correct). They reported that the yoked and random groups showed significant deficits on a later anagram task, relative to the contingent group, but the yoked group performed significantly better than the random group. In addition to contingent, random, and control groups, Benson and Kennelly (1976) used a noncontingent 100% correct group and found no significant differences between the contingent group and the noncontingent/correct group on a later anagram test task, even though the noncontingent/correct group reported having no control over outcomes. Thus, noncontingent positive reinforcement does not appear to produce learned helplessness.

Taken together, these two studies provide evidence that the type and amount of reinforcement directly influence the development of learned helplessness. Similarly, although studies combining noncontingency with a clearly aversive outcome have consistently produced learned helplessness, other studies using the random-reinforcement procedure, which can be seen as alternating noncontingent positive and noncontingent negative reinforcement, have reported mixed results. The overall pattern of results, however, suggests that noncontingent reinforcement is not a necessary and sufficient condition for the development of learned helplessness but that both the contingency and the nature of the obtained outcome are critical to learned helplessness.

In addition to the basic issue of what type of conditions are necessary to produce learned helplessness, there are several other variables that have been shown to exert a significant influence on the development of learned helplessness, including (a) instructional set, (b) task importance, and (c) attributions of performance.

Instructional Set

Instructions given to subjects in the experimental situation regarding response–outcome contingencies have been shown to affect the development of learned helplessness (Geer et al., 1970; Glass & Singer, 1972; Hiroto, 1974). For example, Glass and Singer told subjects they could push a button that would turn off an aversive tone but that the experimenter would "prefer you didn't use it." This procedure significantly increased problem-solving performance during exposure to the tone, relative to those subjects who were not told they had control of the noise. Similarly, Hiroto (1974) reported that sub-

jects who were given instructions that onset of an aversive stimulus was contingent on their responses were significantly less helpless than those subjects who were told the experimenter was controlling the aversive stimulus.

Although not discussed specifically by Seligman, the above results are congruent with the focus of his theory. Since Seligman (1975) posited that behavior is motivated by an internal, cognitive expectancy, these experimental instructions can be viewed as one way of manipulating expectations. For example, Hiroto's (1974) "chance" instructions represent another procedure that induces an expectancy similar to that induced by an experience with an inescapable aversive stimulus (i.e., an expectancy that responses and outcomes are noncontingent), which also results in similar decrements in performance. Similarly, as in the Glass and Singer (1972) study, instructions specifying that responses do control outcomes produce a positive expectancy and no deficits in performance. Thus, instructional set regarding the reinforcement contingencies appears to be a crucial variable for inducing learned helplessness.

Task Importance

Instructions regarding the relative significance of the experimental task have also been shown to influence the development of learned helplessness. As is mentioned above, Seligman (1975) briefly mentioned the effects of task importance but did not elaborate this prediction. Two studies (Roth & Kubal, 1975; Miller & Gold, Note 3) have directly investigated the effect of instructions regarding task importance on the development of learned helplessness. In both studies, importance was manipulated by instructing college student subjects that the training and test tasks were measures of scholastic aptitude and intelligence. Subjects exposed to noncontingent reinforcement showed significantly greater helplessness in the *important* conditions than in the *unimportant* condition. Using contingent, yoked-noncontingent, and random-reinforcement groups, Miller and Gold found no differences between *unimportant* groups but significant differences between all *important* groups. Similar results were reported by Roth and Kubal (1975). These studies strongly support the hypothesis that the perceived importance of the experimental task is a potent factor in the development of learned helplessness and can be manipulated by experimental instructions.

Attributions of Performance

Five studies (Dweck, 1975; Dweck & Reppucci, 1973; Klein et al., 1976; Tennen & Eller, 1977; Wortman, Panciera, Shusterman, & Hibscher, 1976) have investigated the role of the subject's attributions of task performance in the development of learned helplessness. Attribution theory (Weiner, Frieze, Kukla, Reed, Rest, and Rosenbaum, 1971) postulates that the individual's causal attributions influence his or her expectations for probable outcomes of future performance. In learned helplessness, attribution theory would suggest that the subject's attribution concerning the noncontingency of reinforcement would influence both one's expectations and one's performance in future tasks. The research investigating this hypothesis has generally been supportive. Dweck and Reppucci (1973) reported that following experience with noncontingent failure (unsolvable block designs), those children who showed the most performance deficits tended to attribute success or failure to ability; conversely, those children who showed fewest deficits tended to attribute their performance to effort. Dweck (1975) continued this line of research by developing a treatment program for "naturally occurring" helpless children, that is, those who were more adversely affected by failure. In her treatment program, Dweck taught these "helpless" children to attribute failure to lack of effort. Following this "reattribution training," these children showed significant improvements in task persistence and less helplessness than did a group treated with "success-only" experiences, who showed no differences from baseline.

In another study concerning attributions and learned helplessness, Klein et al. (1976) directly manipulated attributions by informing subjects about other subjects' task per-

formance. In the internal-attribution condition subjects were told that 55% of previous subjects succeeded in all problems, and in the external-attribution condition they were told that 90% had failed on all problems. Following these instructions, depressed and nondepressed subjects were exposed to random reinforcement in a discrimination task followed by an anagram test task. Klein et al. reported that the attribution instructions did not significantly affect the nondepressed subjects, but for the depressed subjects the external instructions alleviated learned helplessness on the anagrams. Klein et al. suggested that helplessness and depression are due to both failure and attribution of that failure to personal incompetence.

The results obtained by Tennen and Eller (1977) also support this hypothesis. They exposed subjects to a double helplessness condition, in which subjects were told that each succeeding task was either easier or more difficult. The results indicated that the *easier* group, who presumably made attributions to ability, showed learned helplessness but that the *more difficult* group, who presumably made attributions to task difficulty, did not. Thus, three studies (Dweck & Reppucci, 1973; Klein et al., 1976; Tennen & Eller, 1977) have suggested that attribution of noncontingent failure to ability or personal competence leads to increased learned helplessness, whereas attribution of these outcomes to situational factors or task difficulty does not produce learned helplessness.

One recent study (Wortman et al., 1976) reported findings that appear contradictory to this hypothesis. However, a closer examination of the methodology of this study reveals significant differences between this study and the studies previously described. Wortman et al. told their subjects that their study would deal with the effects of noise on performance and that the amount of noise would be contingent on their performance on several problems. Subjects were then exposed to unavoidable noise and were told they had failed to solve any of the problems. Three information conditions were used: (a) no information, (b) information that another subject could solve the problems (incompetence condition), and (c) information that another subject could not solve the problems (task-difficulty condition). The results indicated that the incompetence group felt more helpless and stressed but performed better than the task-difficulty group did when the same problems were presented later without the noise. Although these results appear contradictory to previous studies, a closer analysis of the methodology clarifies the discrepancy. As was pointed out by Tennen and Eller (1977), the incompetence condition in the Wortman et al. study exposed subjects to a situation in which (a) they were told that the study was about the effects of noise on performance, and (b) another person's performance did not seem to be affected by noise, whereas their own performance under noise conditions was poor. It seems that the most likely attribution of this situation would be that the subject had difficulty solving problems with accompanying noise. Thus, when problems were presented later without noise, the subjects in the incompetence group would expect their performance to improve and would be motivated to attempt to do so. Alternatively, the task-difficulty group, who thought that their failure was due to the difficulty of the task, could not expect any change in their outcomes and thus would have been less motivated to attempt solutions on later problems. Viewed from this perspective, the incompetence group did not produce attributions to the general quality of *competence*, but produced attributions to the relatively specific cause of *ability to solve problems with noise present*, and as such, the incompetence condition was not equivalent to the ability or competence conditions of previous studies.

It is evident from these studies that attributions of noncontingent failure experiences are a potent factor in the development of learned helplessness, with attributions to competence resulting in increased deficits. The implications of these results concerning the effects of attributions are very suggestive and will form the basis of the revised model of learned helplessness proposed in the last section of this article.

More generally, the research reviewed suggests that several variables (task instructions, task importance, attributions) appear to exert significant influences on the development of learned helplessness. Also, these variables appear to be mutually interactive. Since studies have neither varied nor controlled for all of these variables, the relative contribution and specific interactions are difficult to specify at this time.

Facilitation

Although exposure to noncontingent reinforcement has generally been found to result in performance deficits, several studies (Roth & Bootzin, 1974; Roth & Kubal, 1975; Shaban & Welling, cited in Glass & Singer, 1972; Tennen & Eller, 1977; Thornton & Jacobs, 1972; Wortman et al., 1976) have reported improved performance (facilitation) on a later dissimilar task following exposure to a learned helplessness training phase. Since these facilitation effects are at odds with Seligman's basic hypothesis, they deserve closer examination.

Roth and her co-workers (Roth & Bootzin, 1974; Roth & Kubal, 1975) have suggested that these results point to a curvilinear relationship between the amount of exposure to noncontingent reinforcement and learned helplessness. They have proposed that a moderate degree of exposure will result in a greater degree of responding or facilitation, whereas more exposure will result in deficits due to learned helplessness. Seligman (1975) briefly mentioned a similar hypothesis, but only in his argument concerning emotional responses to trauma. He suggested that the organism's initial reaction to uncontrollability is fear, activity, and overresponding, and that continued uncontrollable trauma leads to learned helplessness and depression. Similar predictions were made by Wortman and Brehm (1975) in their integration of reactance theory and learned helplessness.

Since experimental documentation of this hypothesized curvilinear relationship between an amount of exposure and learned helplessness requires at least two levels of exposure to noncontingent reinforcement, only two of the studies reporting facilitation have offered direct evidence concerning this hypothesis. In the first study, Roth and Kubal (1975) varied the amount of exposure to noncontingent outcomes. They reported that subjects exposed to the single helplessness condition showed facilitation effects and that the subjects in the double helplessness condition showed deficits. Analyses of their data also showed significant quadratic trends. These results appear to support Roth and Kubal's hypothesis, but Tennen and Eller (1977) argued that since the instructions of Roth and Kubal's double helplessness condition told the subjects that each succeeding task was "a little bit easier," the Roth and Kubal study confounded the amount of exposure with attribution instructions. Following attribution theory (Weiner et al., 1971), Tennen and Eller hypothesized that attributions to ability (an internal, stable cause) would lead to helplessness, whereas attributions to task difficulty (an external, variable cause) would lead to facilitation. Tennen and Eller tested this hypothesis by replicating Roth and Kubal's study with the addition of a double helplessness condition in which subjects were told that the tasks were becoming more difficult. On a later test task, the double-helplessness "easier" group showed helplessness deficits, and the double-helplessness "more difficult" group showed facilitation effects.

Although the available evidence concerning facilitation effects suggests that a minimum level of exposure to noncontingent outcomes may result in facilitation whereas increased amounts of exposure to noncontingent outcomes may result in learned helplessness, there is evidence that these effects are mediated by the subject's attributions of performance.

In summary, the studies reviewed concerning the etiology of learned helplessness partially support Seligman's hypothesis regarding the necessary and sufficient conditions for the development of learned helplessness. The results of this review indicate that development of learned helplessness requires exposure to environmental conditions in which outcomes are independent of responses *and* are nondesired or aversive to the individual.

The research has also identified four variables (amount of exposure, instructions concerning contingency, task importance, and attributions) that affect the development of learned helplessness and appear to be mutually interactive. Although Seligman's theory briefly discusses the possible effects of instructions and situational importance, it disregards attributions entirely. Furthermore, the focus of the effects of instructions, task importance, and attributions appears to lie in the cognitive processes of the individual, which Seligman does not address beyond the use of a simple expectancy construct. This lack of specification of cognitive processes influencing the development of learned helplessness appears to be a critical deficiency. Finally, a number of studies have reported improved performance following exposure to noncontingent failure outcomes, a phenomenon neither predicted nor explained by Seligman's theory.

Individual Differences

Seligman has not explicitly included any individual differences variables in his formulations of learned helplessness. However, several studies have reported individual differences following exposure to a learned helplessness training phase. Hiroto (1974) reported that subjects scoring in the external direction on Rotter's (1966) Internal–External Locus of Control Scale showed greater performance deficits after exposure to noncontingent reinforcement than did internals. Miller and Seligman (1975), however, did not report differences on this variable. Krantz et al. (1974) investigated the performance of Type A and Type B coronary-prone individuals after exposure to noncontingent reinforcement. They reported significant differences in the responses of Type A and Type B subjects to noncontingent reinforcement. Furthermore, these differences appeared to be a function of the interaction of noncontingency and the intensity of the aversive stimulus, and are consistent with descriptions of these two personality types.

Another individual differences variable that has received some attention in the learned helplessness literature is gender. Although the great majority of learned helplessness studies have not analyzed sex differences, two studies (Dweck & Bush, 1976; Dweck & Reppucci, 1973) have pointed out the possible effects of sex differences in response to noncontingent reinforcement. Dweck and Reppucci (1973) reported that male children attributed outcomes to the amount of effort more often than did female children, and attributions to effort resulted in significantly smaller performance deficits. In a more intensive study, Dweck and Bush exposed male and female children to noncontingent failure from both an adult and a peer evaluator. Failure feedback from adults resulted in impaired performance for girls but in improved performance for boys. Similarly, when a peer evaluator was used, boys showed no performance increases, whereas girls' performance improved significantly. Also, children's attributions of failure varied with the type of evaluator, with boys attributing failure to the amount of effort with the adult evaluator but to their own abilities in the peer situation. Again, girls showed opposite effects, attributing performance to ability with the adult evaluator and to effort with a peer. Although these two studies were done with children and were preliminary, they do offer evidence that there may be sex differences in response to learned helplessness and, furthermore, that attributions of performance may mediate this effect.

Seligman (1974, 1975) has also suggested that reactive depression and learned helplessness are similar disturbances, with common etiology, symptoms, course, and cure. It is not the purpose of this article to discuss the validity of learned helplessness as a model of depression (see Blaney, 1977; Eastman, 1976; Seligman, 1974, 1975), but a brief discussion of the similarities in performance of depressed subjects and subjects exposed to a learned helplessness training phase seems appropriate. Six studies have investigated these similarities and have reported that nondepressed subjects exposed to a learned helplessness training phase perform similarly to depressed subjects on a variety of test tasks, including expectancy changes (Klein & Seligman, 1976; Miller, Seligman, & Kurlander,

1975; Miller & Seligman, 1973, 1976), escape–avoidance learning (Klein & Seligman, 1976), and anagram solution (Klein et al., 1976; Miller & Seligman, 1975). Conversely, exposure to a learned helplessness training phase has not increased performance deficits observed among depressed subjects (Klein et al., 1976; Miller & Seligman, 1975). Thus, Seligman's hypothesis regarding the similarities of depression and learned helplessness has been supported by the available research. It must be noted, however, that all of these studies used nonclinically depressed college students, and the generalizability and relevance of these results for clinical depression remains to be demonstrated.

Although the few studies investigating the effects of individual differences variables and learned helplessness have produced significant results, individual differences, except for depression, have received little attention from learned helplessness research or theory. Since there are both theoretical arguments (Cronbach, 1975; Kiesler, 1971; Underwood, 1975) and empirical evidence (Bowers, 1973; Moos, 1969) suggesting that Person × Situation interactions will provide a more accurate conceptualization of behavior than will either dimension separately, future research and theory in the learned helplessness area should include discussion of these interactions and their effects.

Generalization

Generalization of performance deficits beyond the specific experimental task and situation is a major unresolved issue of the learned helplessness literature. The basic finding in the paradigmatic learned helplessness experiments has been that subjects tend to overgeneralize experiences in the training phase to the later test phase. Subjects exposed to aversive situations in which responses do not influence outcomes overgeneralize this noncontingency to later situations in which responses do influence outcomes. The relative significance of the learned helplessness paradigm is largely tied to the degree of generalization that occurs. Roth and Bootzin (1974) stated the issue succinctly: "The major question is whether an induced external expectancy generalizes to a new situation, not whether it controls behavior in the situation in which it is induced" (p. 255). Clearly, a reduction of responding in a situation in which responses do not influence outcomes is an *adaptive* behavior. It becomes maladaptive only when it is transferred or generalized to new situations in which outcomes are contingent on responses.

Seligman (1975) argued that "men and animals are born generalizers. . . . The learning of helplessness is no exception" (p. 35). He proposed that learned helplessness is a general phenomenon that influences many different aspects of an individual's life. In short, he suggested that learned helplessness can be viewed as a generalized personality trait, which influences behavior in a wide range of situations.

Unfortunately, the evidence concerning the generalizability of deficits caused by learned helplessness inductions is not definitive. Most learned helplessness studies have used test phases in which the type of task and the situational conditions were similar to those of the training phase (Benson & Kennelly, 1976; Dweck & Bush, 1976; Eisenberger et al., 1976; Fosco & Geer, 1971; Hiroto, 1974; Hiroto & Seligman, 1975; Klein & Seligman, 1976; Krantz et al., 1974; Miller & Gold, Note 3). Other studies have demonstrated generalization across types of tasks, when the training and test situations have been similar (Gatchel et al., 1975; Gatchel & Proctor, 1976; Hiroto & Seligman, 1975; Miller & Seligman, 1975; Thornton & Jacobs, 1972). That is, generalization has been reported when subjects perceived that the training and test phases were both part of the same experiment. These studies have not investigated the degree of situational generality of the learned helplessness. Hiroto and Seligman (1975) recognized this shortcoming of their study:

> One limitation on the generality of these effects should be mentioned. The subjects clearly perceived both tasks, as different as they are, as part of the same experiment. We do not know whether any learned helplessness was carried out of the laboratory. (p. 326)

The critical nature of the situational similarity between learned helplessness training and test phases is shown in the Dweck and Reppucci (1973) study, in which children were given a set of unsolvable problems by one teacher and a set of solvable ones by another. Later, these children were administered a similar set of solvable problems by both teachers. The children showed significantly poorer performance when the problems were administered by the teacher who had previously given unsolvable problems. In this study, even though the training and test tasks were identical, the stimulus, namely, the teacher, influenced the children's performance. The children did generalize the experiences of the training phase maladaptively, but the generalization was tied to the situational characteristics in which the learned helplessness was produced.

Other studies have investigated the generalization issue by using a test phase that was situationally dissimilar from the training phase (Roth & Bootzin, 1974; Roth & Kubal, 1975; Sherrod & Downs, 1974; Tennen & Eller, 1977; Wortman et al., 1976; Kilpatrick-Tabak & Roth, Note 4). Of these six studies, three (Roth & Kubal, 1975; Sherrod & Downs, 1974; Tennen & Eller, 1977) reported deficits in their test phase, and three other studies (Roth & Bootzin, 1974; Wortman et al., 1976; Kilpatrick-Tabak & Roth, Note 4) reported no deficits. The studies that did obtain cross-situational generalization had two common factors: (a) prolonged exposure of subjects to noncontingent failure outcomes, and (b) instructions designed to induce an attribution of task performance to ability. Studies that have not exposed subjects to these two conditions have not found cross-situational generalization. Conversely, one study (Kilpatrick-Tabak & Roth, Note 4) exposed subjects to these conditions and did not obtain generalization. Thus, cross-situational generalization of learned helplessness has not been conclusively demonstrated at this time.

In summary, there is evidence that learned helplessness generalizes from one type of task to another, but there is no conclusive evidence regarding the degree of generalization across situations, and no study has varied both situational and task dimensions. The failure to demonstrate cross-situational generalization is a major flaw in the learned helplessness literature. Without conclusive evidence for generalization, the significance of the learned helplessness phenomenon becomes questionable. Unfortunately, Seligman's theory does not address the issue of generalization in any detail and does not adequately specify the determinants of generalization of learned helplessness. If learned helplessness is to represent an adequate analogue of depression, then a detailed statement of how and when generalization will occur and relevant research are essential.

Alleviation

Drawing on the animal research, Seligman (1975) suggested that learned helplessness could be "cured" by the establishment of an expectancy that outcomes are dependent on responses. In the case of dogs in the shuttle box, this "therapy" consisted of repeatedly pulling dogs to the safe side of the box until they learned that their movement affected the offset of the shock. Four studies have attempted to alleviate deficits caused by learned helplessness inductions in humans. Of these, three (Dweck, 1975; Klein & Seligman, 1976; Kilpatrick-Tabak & Roth, Note 4) employed procedures in which subjects were given response-dependent feedback concerning their performance on a task interposed between training and test tasks. Klein and Seligman (1976) exposed nondepressed subjects to inescapable noise and then exposed half of these "helpless" subjects and depressed subjects without pretraining to a treatment of either 4 or 12 discrimination problems with response-dependent feedback. They reported that on a subsequent escape–avoidance task, both treatment groups showed significantly better performance and a greater expectancy of response–outcome dependence than did subjects in the no-treatment groups. This study closely paralleled the earlier research with dogs and supported Seligman's hypothesis that exposure to response-dependent outcomes reduces learned helplessness.

A second study that attempted to alleviate learned helplessness was done by Kilpatrick-Tabak and Roth (Note 4). In this study, subjects selected without respect to depression were exposed to a training phase used by Roth and Kubal (1975). These subjects and depressed subjects without pretraining were then exposed to one of four treatments: (a) reading Velten's (1968) list of positive self-statements, a procedure that may be broadly conceptualized as a cognitive, possibly reattributional task; (b) solution of a set of simple anagrams; (c) waiting alone for 15 minutes; and (d) waiting with another person for the same time. All subjects, including a nondepressed control group without training or treatment exposure, were given the test task from Roth and Kubal's (1975) study. The results indicated that although both Treatments 1 and 2 were successful for the nondepressed, helpless group, they did not alleviate the deficits shown by the depressed subjects. In fact, exposure to simple anagrams increased later deficits in depressed subjects, a result quite contrary to Seligman's hypothesis. Although this study was compromised by the fact that the waiting-period groups did not differ from the control group on the test task, the failure of the treatment procedure (i.e., solving anagrams) to alleviate deficits in the depressed subjects conflicts with Seligman's hypothesis and the results of Klein and Seligman (1976). Thus, the evidence regarding Seligman's hypothesis is inconclusive at this time.

Two other studies (Dweck, 1975; Klein et al., 1976) have taken a different approach to alleviation of learned helplessness and have focused on the subjects' attributions of learned helplessness performance. Dweck (1975) selected children who were identified as "helpless" by several school personnel. She then exposed these children to one of two treatment conditions. The first treatment, "reattribution training," taught the children to attribute failure to a lack of effort. The second treatment, "success only," provided children with a variety of success experiences with no attribution training. The results indicated that the success-only group did not change from their baseline level of performance, whereas the reattribution group showed significant improvements from their baseline of performance. Unfortunately, the reattribution-training group included contingent practice, so conclusions of differential effectiveness cannot be attributed solely to the cognitive intervention. The most that can be concluded is that a package of response-dependent feedback and altered cognitions is superior to response-dependent feedback alone.[2]

Klein et al. (1976) reported that the performance of depressed subjects improved when they were told that most people failed the training task, an instructional set presumed to induce an attribution of task difficulty. However, the treatment used by Klein et al. did not improve the performance of nondepressed subjects exposed to response-independent failure outcomes.

In summary, the literature pertaining to the alleviation of learned helplessness provides few clear conclusions. Naturally occurring performance deficits found in depressed college students and children have been improved by reattribution therapy (Dweck, 1975; Klein et al., 1976) and, in one study, by exposure to response-dependent success conditions (Klein & Seligman, 1976). Another study, however, found that exposing depressed subjects to response-dependent success outcomes produced no improvement (Kilpatrick-Tabak & Roth, Note 4). On the other hand, deficits engendered in nondepressed college students have been alleviated successfully in two studies (Klein & Seligman, 1976; Kilpatrick-Tabak & Roth, Note 4) by exposure to response-dependent feedback, but Klein et al. (1976) reported that treating nondepressed college students with a reattribution treatment was not successful in alleviating the produced deficits. The two investigations that have compared response-dependent success and a cognitive manipulation either did not afford unconfounded interpretation (Dweck, 1975) or did not include a treatment focused on reattributing failure experiences (Kilpatrick-Tabak & Roth, Note 4). Thus,

---

[2] The authors are indebted to Larry Young for this observation.

present research does not allow conclusions concerning the relative effectiveness of these treatments.

Summary

As the preceding review demonstrates, Seligman's theory of learned helplessness no longer offers a full and viable explanation for the results of the current research. There are several areas in which Seligman's model seems inadequate to account for present data. The first area concerns the necessary and sufficient conditions for the development of learned helplessness. Although Seligman hypothesized that an expectancy of response–outcome independence is necessary and sufficient for learned helplessness to occur, current research suggests that an expectancy of response–outcome independence and a nondesired outcome are necessary for the development of learned helplessness. The second area in which Seligman's theory seems deficient is in regard to the other factors that effect learned helplessness. More specification is needed to explain how, why, and when variables such as instructions, task importance, and subject's attributions effect learned helplessness. The third area concerns generalization. Seligman's theory does not delineate the processes and circumstances under which learned helplessness will generalize. Since the relevance of the learned helplessness paradigm seems to hinge on generalization and present research has not documented generalization of learned helplessness, this is a most critical deficiency. The lack of integration of individual differences factors into learned helplessness theory is a fourth area of weakness. The final area concerns the alleviation of learned helplessness. Support for Seligman's hypothesis concerning the conditions necessary for alleviation of learned helplessness is equivocal, with attributions appearing to be a major factor.

More generally, research and theory in the learned helplessness area appear to be plagued by a narrowness of approach. First, the basic stance of Seligman's theory has largely been neglected by researchers. Seligman hypothesized that the cognitive expectancy of response–outcome independence, not the actual experience of exposure to those conditions, produces learned helplessness. Unfortunately, the learned helplessness research literature has focused primarily on the environmental conditions that seem to produce learned helplessness and have neglected the cognitive schema that Seligman postulated as crucial. Seligman's theory can be seen as contributing to this neglect because it does not elaborate the cognitive processes and variables relevant to learned helplessness. One reason for this deficiency may lie in the animal research origins of the learned helplessness paradigms. Dogs and rats simply do not have the cognitive complexity or construction abilities of humans. As was pointed out by Levis (1976), Seligman's theory equates the cognitive processes of humans with those of a cockroach. This article takes the position that the explanation of human behavor requires more complex and detailed hypotheses regarding the function of cognitive processes. The learned helplessness research appears to support this contention. Recent research has pointed to the importance of cognitive processes in mediating the development of learned helplessness, and further research and theory must focus on these cognitive aspects of learned helplessness.

It is apparent that learned helplessness has been conceptualized primarily as a *situational* paradigm; that is, researchers have focused on delineating the situational conditions that produce learned helplessness in a majority of subjects. In doing so, however, the literature has largely neglected the processes by which these situational experiences are translated into cognitive schemata, the characteristics of these schemata, and the processes by which cognitive schemata influence future behavior. It is our contention that the situational view of human behavior does not adequately represent or explain learned helplessness.

In view of the deficiencies in the present theory of learned helplessness noted above, the possible significance of the learned helplessness phenomenon, and the continued interest and research in this area, it seems appropriate at this time to propose a revised model of human learned helplessness. The model presented below attempts to outline a

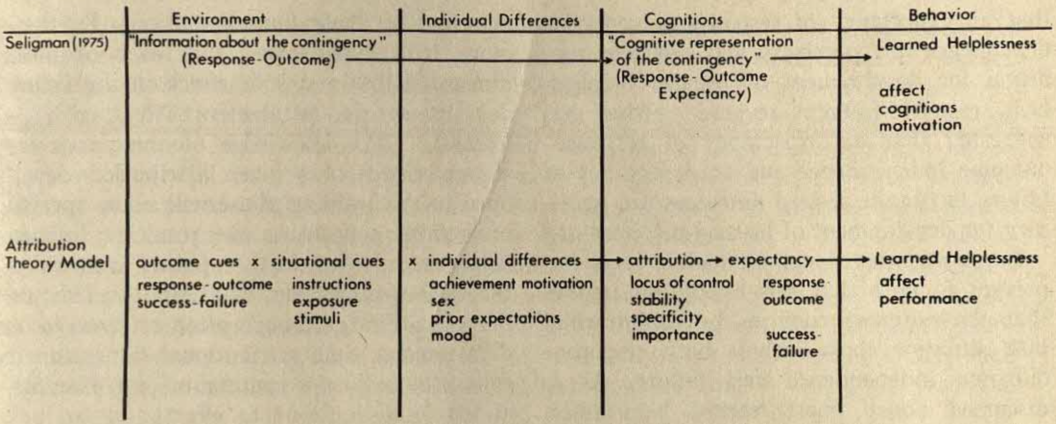

Figure 1. A comparison of theories of learned helplessness.

modified version of Seligman's theory of learned helplessness that better fits the available data and that gives definitive directions for new research in the area.

Attribution-Theory Model

Figure 1 presents the basic framework of both Seligman's model and the present revised model of learned helplessness. The present model differs from Seligman's model in several general areas. The first difference concerns the complexity of the antecedents of learned helplessness. Seligman (1975) postulated a single term of "information about the contingency" (p. 49), which refers to the environmental cues concerning response–outcome dependency, whereas the present model subdivides environmental events into two categories: outcome cues and situational cues. Outcome cues refer to the characteristics of the feedback concerning an individual's performance in a given situation, including Seligman's concern with response–outcome dependency. Situational cues refer to the stimuli present in the situation itself that influence the individual's perception and interpretation of outcomes. Additionally, the revised model includes an individual differences term and is explicitly interactional. Environmental events of outcomes and situational cues are hypothesized to interact with each other and with individual differences variables to influence the individual's cognitive processes. The final difference concerns the cognitive processes involved in learned helplessness. In addition to Seligman's expectancy term, the present model includes an attribution term.

In keeping with the previous focus on the five salient issues concerning learned helplessness, these variables and the present model are discussed below in more detail within the context of these issues.

Nature

In the present model, the nature of learned helplessness is presumed to be similar to that originally proposed by Seligman (1975). It differs from the earlier model by separating deficits resulting from learned helplessness into affective and performance components rather than into affective, motivational, and cognitive aspects. The binary distinction conforms more closely with current research, which has not successfully isolated cognitive and motivational deficits. The present model also differs from Seligman's position by suggesting that the affective and performance components of learned helplessness can occur independently and will vary with the content of the individual's attributions.

Etiology

The revised model makes two major additions to Seligman's theory concerning the etiology of learned helplessness. The first change concerns the content of expectancy that is hypothesized to produce learned helplessness. Although Seligman (1975) suggested

that an expectancy of response–outcome independence is a necessary and sufficient condition for development of learned helplessness, recent research, reviewed above, has suggested that an expectancy of response–outcome independence and an expectancy of failure to obtain desired outcomes are necessary for development of learned helplessness.[3]

From an operational viewpoint, it is important to note that this hypothesis implies that the outcome cues in the experimental task situation must include both response–outcome independence *and* failure. As is discussed above, most learned helplessness studies have used procedures that expose subjects to failure as well as noncontingent outcomes. In the remainder of this article, outcomes that are response independent and undesired are referred to as *learned helplessness outcomes.*

The second addition concerning etiology in the revised model is the inclusion of the attribution term, which is derived primarily from the theory and research of attribution theory (Weiner, 1974; Weiner et al., 1971). In contrast with the single expectancy term of Seligman's model, attribution theory suggests that analysis of the individual's ascriptions of causality of environmental events will lead to more accurate representations of cognitive processing and to better prediction of future behavior. Seligman suggested that exposure to environmental events leads to the development of an expectancy concerning control and that this expectancy in turn influences future behavior, whereas an attribution model suggests that the interaction of outcome and situational cues with individual differences variables results in an attribution to explain learned helplessness outcomes, and the characteristics of this cause then determine the expectancy that influences future behavior.

According to Weiner (1974), attributions can be characterized by two basic dimensions: locus of control (internal vs. external) and stability (stable vs. variable). The present model adds two additional dimensions that seem particularly relevant to learned helplessness: specificity (specific vs. general) and importance (important vs. unimportant). It is predicted that any attribution can be characterized by these four dimensions. Furthermore, it is hypothesized that each of these dimensions has a specific effect on the future development and parameters of learned helplessness. Thus, knowledge of the particular characteristics of a given attribution developed in the training phase will allow specific predictions concerning the resulting learned helplessness. Two further points concerning the various dimensions of attributions should be noted. First, although often referred to as dichotomous, each attributional dimension is conceptualized as a continuum; for example, although attributions to effort and to luck may be conceptualized as variable, luck may be more variable than effort. Second, each dimension of attribution is a subjective dimension; that is, one individual may perceive luck as more stable ("I'm a lucky person") than another ("That was just luck"). Thus, although general statements can be made concerning the dimensions of particular attributions, for best prediction the individual's perception of each dimension is necessary.

Locus of Control

The locus of control dimension represents a concern similar to the response–outcome contingency that Seligman postulated as basic to learned helplessness. However, the locus of control term used in this model does not represent the contingency of environmental events or an expectancy of response–outcome contingency but instead represents the assignment of causality for the perceived contingencies to internal or external sources. For example, in a typical learned helplessness study, subjects may perceive that their responses do not influence outcomes but may assign causality for this noncontingency to

---

[3] It should be noted that the present model treats task outcome and response–outcome contingency as orthogonal, independent dimensions. Although there is some evidence that these dimensions may not be independent (Jenkins & Ward, 1965), the relationship between outcome and contingency is complex and poorly understood at present (see Blaney, 1977). It seems prudent to treat these dimensions as independent and to use separate measures and predictions for each variable, until future research clarifies the relationship.

an internal source ("I am stupid") or an external source ("The experimenter is controlling the task"). Research in attribution theory (Eswara, 1972; Rest, Nierenberg, Weiner, & Heckhausen, 1973; Weiner & Kukla, 1970) has suggested that the attributional dimension of locus of control directly influences the subjects' affective reactions to task performance. These studies support the hypothesis that if one attributes failure to an internal cause, self-depreciation and negative affect result, whereas attribution of failure to an external cause minimizes this affect. Similarly, positive affects following success are maximized by attributions to internal causes and are minimized by attributions to external causes. These findings suggest that following exposure to learned helplessness outcomes, an attribution to an internal cause will produce negative affect, and an attribution to an external cause will reduce this negative affect.

Evidence for this hypothesis comes from the Roth and Kubal (1975) study, in which half the subjects were exposed to a manipulation designed to produce a belief that the learned helplessness training task was a measure of intelligence. This *important* group reported significantly more depression and frustration following training than did an *unimportant* group. If one assumes that task performance in the *important* condition was attributed to intellectual ability, an internal cause, then this study can be seen as supporting this hypothesis. Unfortunately, this study and others in the learned helplessness area have not focused on attributions and have not differentiated among the dimensions proposed in the present model. That is, although the Roth and Kubal manipulation can be hypothesized to produce an attribution to intellectual ability, no direct measures of attributions were taken, and furthermore, an attribution of intellectual ability can be characterized by other dimensions than locus of control, which may have influenced emotional responses. Thus, the Roth and Kubal study offers suggestive but nonconclusive support for the present model. Unfortunately, the presence of suggestive but nonconclusive evidence will be a common finding for most of the hypotheses of the model. The research focusing on the relationship of these dimensions of attributions to learned helplessness simply has not been done.

This hypothesis also offers an explanation for those studies that have reported an increase in self-reported anger and hostility following exposure to learned helplessness training conditions. As is discussed above, these results are not explained by Seligman's model. However, the attribution theory model suggests that an attribution of learned helplessness outcomes to an external source, such as experimenter control, would be predicted to reduce depression but also could be expected to increase the subject's anger at the external agent. Anecdotal evidence for this hypothesis comes from Miller and Seligman (1975, 1976), who stated that two subjects in each study reported high levels of anger after the training phase and told the experimenter that "they had decided early in the experiment that it had been rigged so that they could not escape the noise and that this had made them angry" (Miller & Seligman, 1975, p. 235). Moreover, these subjects reported large decreases in depression. This anecdotal evidence provides further support for the hypothesis that attributions mediate affective reactions to learned helplessness outcomes.

Stability

Stability refers to the relative permanence associated with an attribution. Environmental events can be attributed to causes that are stable or variable. For example, *intelligence* is a relatively stable attribution, whereas *effort* or *luck* are relatively variable. Attribution theory hypothesizes that stability of attribution determines the degree of influence that past outcomes exert on expectancies for performance in future situations. Several studies (Fontaine, 1974; McMahan, 1973; Weiner et al., 1976) have supported this hypothesis. Thus, if one attributes past outcomes to luck (a variable cause) then these outcomes will not influence one's expectancies in future situations, but if one attributes past outcomes to ability (a stable outcome) then one's expectancies for performance in future situations will shift in the

direction of the outcome. These results suggest that stability of attribution mediates the degree of influence that past outcomes exert on expectancies for performance in future situations, or the degree of cross-situational generalization.

In the learned helplessness situation, in which the outcome cues to be explained are those of response–outcome independence and failure, this hypothesis suggests that an attribution to a stable cause will result in a shift of the subjects' expectancies for future performance toward noncontingency and failure and, therefore, generalization to dissimilar situations. This hypothesis, then, goes right to the heart of the learned helplessness paradigm, suggesting that the stability of attribution predicts the situational generalization of learned helplessness. Thus, the next postulate of the model suggests that an attribution of learned helplessness outcomes to a stable cause will tend to increase the situational generalization of these outcomes.

This hypothesis concerning cross-situational generalization is supported by the Dweck and Reppucci (1973) study in which the greatest cross-situational generalization of learned helplessness deficits occurred in those children who attributed performance to ability, a stable cause, and children who showed the fewest deficits attributed their performance to effort, a variable cause. The results of Roth and Kubal (1975) and Tennen and Eller (1977) also lend support to this hypothesis. In these studies cross-situational generalization was found, but only in the experimental groups who had been exposed to an experimental manipulation designed to induce attributions to intellectual ability, a stable cause.

Specificity

Although attribution theory does not address itself to this dimension, attributions can also be characterized by their specificity or generalizability. In the learned helplessness paradigm, the specificity of attribution is postulated to predict the number of future tasks that will be affected by the expectancies developed in the training phase. For example, if a subject attributed learned helplessness outcomes to the ability to solve discrimination problems, a specific cause, then expectancies for and performance on future math problems should not be affected. But if the attribution were to a general cause, ability to solve problems, then expectancies for and performance on future math problems should be affected. Thus, the present model hypothesizes that attribution of learned helplessness outcomes to a general cause will tend to increase the influence of the training-phase outcomes on expectancies and performance in different types of tasks. Unfortunately, no study has directly investigated the effects of this dimension on learned helplessness.

Importance

The last dimension of attribution to be discussed is subjective importance, that is, the relative value a person assigns to an event. Subjective importance or reinforcement value has been discussed as a major component of the social-learning theory of human behavior (Rotter, 1954; Rotter, Chance, & Phares, 1972) but has not been discussed in detail by attribution theorists. In social-learning theory, the construct of subjective importance is seen as a major dimension of the task. The present model, however, in keeping with an emphasis on attributions, predicts that attributions, as well as tasks, differ in the degree of importance to a given individual and that the subjective importance of an attribution can also influence present and future behavior. This model further suggests that the influence of the dimension of subjective importance will be manifested primarily in terms of the magnitude of the affective and performance deficits predicted by the other dimensions of attributions.[4] Thus, it is hy-

[4] The subjective importance of the attribution is not the only factor that contributes to the magnitude of deficits. As social-learning theory and Abramson, Seligman, and Teasdale (1978) have mentioned, the subjective importance of the outcome itself can also influence magnitude. Similarly, the other dimensions of attributions will affect magnitude. Attributions to causes that are relatively stable or general will, due to their relative generalizability, increase the magnitude of learned helplessness deficits.

pothesized that the magnitude of learned helplessness deficits will covary with the subjective importance of the attribution of learned helplessness outcomes.

Two studies designed to vary subjective importance (Roth & Kubal, 1975; Miller & Gold, Note 3) have manipulated experimental task cues to increase the subjects' belief that task performance reflected intellectual ability. Although there are no direct data, this type of manipulation appears to increase the probability of an attribution to intellectual ability, presumably an important attribution for most college students. Roth and Kubal reported that subjects exposed to the noncontingent *important* manipulation reported increased depression and anxiety and increased performance deficits. Miller and Gold also reported significantly greater performance deficits in their noncontingent *important* condition. Of course, an attribution of intellectual ability can also be characterized as internal, stable, and general, so it is not clear that the importance of the attribution is *the* causal factor.

The basic etiological hypotheses of the revised model of learned helplessness have now been presented. Exposure to outcome cues of response–outcome independence and failure results in an attribution to explain these outcomes. The characteristics of the attributions constructed are predicted to determine the development, type, and generalization of learned helplessness deficits. The dimension of locus of control is predicted to determine the resulting affective components of learned helplessness, with an attribution of learned helplessness outcomes to an internal cause hypothesized to result in depression and anxiety. Stability of attribution is predicted to determine cross-situational generalization, with attributions to a stable cause resulting in cross-situational generalization of learned helplessness deficits. Specificity of attribution is hypothesized to control the cross-task generalization, with attributions to general causes resulting in cross-task generalization. The importance dimension is hypothesized to affect the intensity of the deficits produced in learned helplessness, with attributions to important causes resulting in maximum disability. Combining these hypotheses, it should be noted that attributions to causes that are internal, important, stable, and general are predicted to maximize the severity and generalization of learned helplessness, whereas attributions to causes that are external, unimportant, variable, and specific will minimize deficits.

As with other cognitive processes, the construction of attributions can be influenced by a variety of other factors. Thus, the next task of the model is to specify those situational cues that may interact with outcome cues and individual differences variables to affect the attributional process.

Situational Cues

As is mentioned above, situational cues refer to the stimuli present in the situation that exert influence on the subjects' attributions of task outcome. Several situational task cues have been identified in the learned helplessness literature, including (a) instructions regarding the contingency of responses and outcomes (Geer et al., 1970; Glass & Singer, 1972; Hiroto, 1974), (b) instructions regarding task difficulty (Klein et al., 1976), (c) instructions and other task cues specifying the nature of the task (Roth & Kubal, 1975; Miller & Gold, Note 3), and (d) the amount of exposure to learned helplessness conditions (Roth & Kubal, 1975). Other situational cues that are hypothesized by attribution theory to influence the development of attributions include social norms, observation of others' performance, and type of task (Weiner, 1974).

Space limitations prevent a discussion of the specific effects of all of these situational cues (see Weiner, 1974), but an example may prove helpful in understanding the significance of this class of variables. Klein et al. (1976) reported that instructions that the task was difficult tended to increase attributions of learned helplessness outcomes to task difficulty, an external, variable, and specific cause, whereas instructions that the task was easy tended to increase the probability of attributions to ability, an internal, stable, and general cause. These instructions also produced corresponding differences in task performance. Thus, it can be seen that

situational cues interact with outcome cues to produce an attribution that then influences expectancies for future performance and later task behavior.

Individual Differences Variables

Individual Differences Variables is perhaps a misnomer for this section, since a majority of the variables discussed are interactions between situational and individual differences variables. However, the purpose of this section is to specify those individual differences variables that may interact with situational variables to produce differential influences on learned helplessness performance. Since a number of individual differences variables have been found to influence the attributional process (see Weiner, 1974), this article presents only those variables specifically relevant to learned helplessness.

One individual differences variable that has been shown to influence attribution is gender. As is discussed above, Dweck and Reppucci (1973) reported different attributions from male and female children, with males tending to attribute failure outcomes to effort and females tending to attribute them to ability. These results suggest that females would tend to be more susceptible to learned helplessness manipulations. However, as another study (Dweck & Bush, 1976) reported, these results appear to be further mediated by the situational cue of the role of the evaluator (peer vs. adult). Clearly, then, sex is a potent variable that can influence attributions and development of learned helplessness. It is also clear that the interaction among sex, learned helplessness, and attributions is a complex one, with other variables playing a major role. These results also indicate that sex is one variable that should be controlled in future learned helplessness research.

Another individual differences variable that has been shown to influence the development of attribution is prior expectancies of future outcomes. Several studies have reported data that suggest that the congruency between prior expectancies and present outcomes can influence the attributional process. Feather (1969; Feather & Simon, 1971a, 1971b) assessed subjects' prior expectancies by subjects' reports (Feather, 1969) or previous experience (Feather & Simon, 1971a, 1971b). Subjects were then given anagram problems and were asked to rate the causes of performance outcomes. These results suggested that an outcome that is discrepant from previously held expectancies tends to be attributed to external causes, and outcomes congruent with expectancies tend to be attributed to internal causes. These results were replicated by Gilmor and Minton (1974). In a study investigating another dimension of attributions, McMahan (1973) also used the anagram problems and found that expectancy disconfirmation produced attributions to variable causes and that expectancy confirmation led to attributions to stable causes.

These results suggest that individuals tend to attribute discrepant outcomes to causes that lead to the smallest change in the individual's previously held expectancies. If this is true, then one would also predict that discrepant outcomes would be attributed to specific causes rather than to general causes. Unfortunately, there has been no reported research bearing directly on this question. However, the available research does suggest that when learned helplessness outcomes are highly discrepant from previously held expectancies, attributions tend to be made to external, variable, and specific causes. Conversely, if learned helplessness outcomes are congruent with previously held expectancies, attributions tend toward internal, stable, and general causes. Thus, learned helplessness is predicted to be more likely to occur when training-task outcomes are congruent with prior expectancies than when outcomes are incongruent. This prediction is supported by Hiroto's (1974) study, in which subjects with prior expectancies of an external locus of control (noncontingent) showed greater deficits following exposure to a noncontingent failure experience than did subjects with a belief in an internal locus of control.

A final important individual differences variable for the learned helplessness paradigm is depression. Since learned helplessness has been advanced as a theory of depression, the interaction between depressed mood and exposure to learned helplessness outcomes

should be explored in some detail. Depressed subjects have been shown to be generally more pessimistic about future outcomes than are nondepressed subjects (Beck, 1974; Loeb, Beck, & Diggory, 1971)[5] and to expect the independence of responses and outcomes (Calhoun, Cheney, & Dawes, 1974). Since learned helplessness outcomes of noncontingency and failure are congruent with these expectations of depressed subjects, the present model suggests that depressed subjects would tend to attribute learned helplessness outcomes to causes that are internal, stable, and general. Current research has generally supported this prediction. Klein et al. (1976) reported that depressed subjects tended to attribute performance to ability, an internal, stable, and general cause, after failure but not after success. Similarly, Miller and Seligman (1973) reported that depressed subjects tended to show smaller expectancy changes than did nondepressed subjects on a skill task after failure, but there was no difference after success. Since a small expectancy change has been shown to be associated with a stable attribution (Weiner et al., 1976), Miller and Seligman's results suggest that depressed subjects tend to attribute noncontingent failure on skill tasks to stable causes.

If depressed subjects tend to attribute learned helplessness outcomes to internal, stable, and general causes, then the present model predicts that depressed subjects would exhibit greater performance deficits and greater affective reactions than would nondepressed subjects, following exposure to learned helplessness outcomes. Two studies (Hammen & Krantz, 1976; Wener & Rehm, 1975) offer unambiguous support for this hypothesis. In both studies, depressed and nondepressed subjects were exposed to noncontingent success and failure outcomes. Following failure, depressed subjects' responses showed more depression, less self-confidence, and decreased expectancies for future success. No difference between groups was reported with success outcomes. Although these results support the hypothesis of this model, they appear inconsistent with the learned helplessness studies of Miller and Seligman (1975, 1976), which reported that depressed subjects exposed to learned helplessness training do not differ from depressed subjects without learned helplessness exposure. However, as is discussed above, Miller and Seligman (1975, 1976) reported that in both studies, two of eight depressed subjects exhibited performance similar to that of nondepressed/no learned helplessness subjects on the test task. The present view is that these four subjects may have attributed learned helplessness outcomes to experimenter control, an external, variable, and specific cause. Thus, for these depressed subjects, although learned helplessness outcomes may have been congruent with their prior expectancies, other situational cues may have led to attributions of experimenter control, which in turn may have led to nondepressed performance and therefore to a reduction in the difference between the depressed/learned helplessness group and the depressed/no-treatment group. Miller and Seligman (1976) reported that the remaining depressed/learned helplessness subjects did show greater performance deficits than the depressed/no-treatment group, but these differences were not statistically significant. They further suggested the possible operation of a floor effect not found in the Hammen and Krantz (1976) or Wener and Rehm (1975) studies. These data, though somewhat ambiguous, tend to support the present hypothesis, that depressed subjects tend to attribute outcomes of noncontingency and failure to internal, stable, and general causes and therefore tend to exhibit greater depression and performance deficits.

This hypothesis, in combination with the basic etiological hypotheses of the model, suggests the following chronology of reactive depression: Due to some combination of situational cues and repeated exposure to noncontingent and nondesired outcomes, the individual's attributions of these outcomes

---

[5] Miller and Seligman (1973) and Hammen and Krantz (1976) have reported that depressed subjects do not differ from nondepressed ones in initial task expectancies. These studies, however, only used nonclinically depressed subjects and assessed only specific task expectancy. The lowered expectancy discussed above reflects a more generalized expectancy, which is probably more pronounced in clinically depressed populations.

changes from external, variable, and specific causes to internal, stable, and general causes. This change results in a change in future expectancies, performance, and mood. Thus, in new situations, the individual expects noncontingency and failure, and when these congruent outcomes occur, they are attributed to internal, stable, and general causes, whereas discrepant outcomes of success and contingency are attributed to external, variable, and specific causes and do not influence future expectancies, performance, or mood. The individual is then depressed and tends to disregard outcomes of success and contingency while overgeneralizing failure and noncontingent outcomes. Thus, response initiation declines, a greater number of failure and noncontingent outcomes do occur, and the vicious circle of depression has begun.

Generalization

As is reported above, the lack of cross-situational generalization is a major obstacle to the significance of the learned helplessness paradigm. Seligman's theory offers no convincing explanation for the lack of results. As is suggested above, in the Etiology section, the present revised model hypothesizes that cross-situational generalization of learned helplessness is a function of the stability dimension of the attributions constructed. According to this hypothesis, cross-situational generalization should result if and when attributions of learned helplessness outcomes are made to relatively stable causes and should not result when attributions are made to relatively variable causes.

However, performance deficits in a situationally similar task may be due to attributions that are relatively variable but that are stable for the duration of the experiment, such as experimenter control or task difficulty, or such deficits may be due to relatively stable causes such as ability or personality. The present model suggests that performance deficits that occur as a result of attributions to relatively variable causes are really "pseudohelplessness" because they do not represent a change in the individual's basic expectancies or mode of adaptation. Deficits that occur as a result of attributions to stable causes do result in more lasting, generalized learned helplessness.

Awareness of the role of stability of attributions in the generalization of learned helplessness leads to several other hypotheses and explanations concerning generalization of learned helplessness. First, since the typical nondepressed college student tends to expect response–outcome dependence and success (Parducci, 1963, 1965, 1968), the outcomes of the learned helplessness training phase will be highly discrepant and thus are predicted to be attributed to causes that are relatively variable. Since attributions to variable causes are not predicted to result in cross-situational generalization of deficits, the present model predicts that studies using nondepressed college students would produce temporary, specific pseudohelplessness deficits during the experimental task situation, but not the relatively permanent, generalizable learned helplessness described by Seligman, unless very strong situational cues designed to produce more stable attributions were utilized. One of the few studies to report cross-situational generalization used such situational cues. Roth and Kubal (1975) used three situational manipulations, each of which would tend to increase the probability of an attribution to a stable cause and thus to cross-situational generalization of learned helplessness.

This interpretation of the learned helplessness research is crucial to future research in the area because (a) it explains the current lack of documented cross-situational generalization, and (b) it suggests that to produce generalized learned helplessness, the experimental situational cues will have to be manipulated to induce attributions to relatively stable causes.

Alleviation

The final statement of the present model concerns alleviation or treatment of learned helplessness. Following the basic statement of the model, treatment of learned helplessness is focused on changing the subjects' attributions. The first step in treating learned helplessness is assessing the type of learned helplessness involved. Are the obtained defi-

cits due to the pseudohelplessness produced by an attribution to external, variable, and specific causes, or are deficits due to learned helplessness produced by internal, stable, and general attributions? If the deficits are due to pseudohelplessness, then a change in outcomes or situations would be sufficient to change future expectancies and performance. That is, if one attributed learned helplessness outcomes of response–outcome independence and failure to a variable or specific cause, then a change in these outcomes will be sufficient to change expectancies for future outcomes. This reasoning suggests that for pseudohelplessness, exposure to experiences of response–outcome dependence and success will alleviate deficits. The success of this treatment with nondepressed subjects has been demonstrated by Klein and Seligman (1976) and Kilpatrick-Tabak and Roth (Note 4).

On the other hand, if deficits are due to learned helplessness or depression caused by an attribution of learned helplessness outcomes to internal, stable, and general causes, then exposure to response-dependent success will be attributed to variable and specific causes and will not influence future expectancies of performance. The failure of exposure to response-dependent success to alleviate in depressed subjects was demonstrated by Kilpatrick-Tabak and Roth (Note 4). According to the present model, treatment of learned helplessness or depression would require a direct focus on changing the attributions themselves. Examples of these changes and the effectiveness of this type of treatment can be seen in the studies of Klein et al. (1976) and Dweck (1975).

The revised formulation of learned helplessness has now been presented. The interaction of situational cues, outcome cues, and individual differences was hypothesized to result in an attribution to explain learned helplessness outcomes. The characteristics of this attribution were postulated to mediate the influence of learned helplessness outcomes on expectancies and behavior in future situations. Four dimensions of attributions were discussed: (a) locus of control, (b) stability, (c) specificity, and (d) importance. Attributions characterized as internal, stable, general, and important were predicted to maximize learned helplessness.

This model appears to offer a more accurate and predictive theory of learned helplessness. We hope that the model will open new avenues of research within the learned helplessness paradigm. Beyond testing the specific hypotheses of the model itself, the focus on cognitive processes and attributions may lead to an increased awareness of the complexity of human cognition and the inclusion of measures of cognition in experimental designs. Also, the present model can serve as a springboard for investigation of interactions of outcome and situational cues with individual differences variables and of the relations of these interactions with cognitive processes and future behavior.

Reference Notes

1. Seligman, M. E. P. *Depression and learned helplessness: Recent developments.* Paper presented at the meeting of the Association for the Advancement of Behavior Therapy, New York, December 1976.
2. Seligman, M. E. P. Personal communication, August 11, 1977.
3. Miller, I. W., III, & Gold, J. *The relationship of percentage of reinforcement and situational importance in the development of learned helplessness.* Manuscript submitted for publication, 1978.
4. Kilpatrick-Tabak, B., & Roth, S. *An attempt to reverse performance deficits associated with depression and experimentally induced learned helplessness.* Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1976.

References

Abramson, L. Y., Seligman, M. E. P., & Teasdale, J. D. Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, 1978, *87*, 49–74.

Amsel, A. Behavioral habituation, counter-conditioning, and a general theory of persistence. In A. Black & W. Prokasy (Eds.), *Classical conditioning II: Current theory and research.* New York: Appleton-Century-Crofts, 1972.

Beck, A. T. The development of depression. In R. J. Friedman & M. M. Kurtz (Eds.), *The psychology of depression: Contemporary research and theory.* New York: Wiley, 1974.

Benson, J. S., & Kennelly, K. J. Learned helplessness: The result of uncontrollable reinforcements or uncontrollable aversive stimuli? *Journal of Personality and Social Psychology*, 1976, *34*, 138–145.

Blaney, P. H. Contemporary theories of depression: Critique and comparison. *Journal of Abnormal Psychology*, 1977, *86*, 203–223.

Bowers, K. Situationism in psychology: An analysis and critique. *Psychological Bulletin*, 1973, *80*, 307–336.

Braud, W., Wepman, B., & Russo, D. Task and species generality of the "helplessness" phenomenon. *Psychonomic Science*, 1969, *16*, 154–155.

Calhoun, L. G., Cheney, T., & Dawes, A. S. Locus of control, self-reported depression, and perceived causes of depression. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 736.

Cronbach, L. J. Beyond the two disciplines of scientific psychology. *American Psychologist*, 1975, *30*, 116–126.

Dweck, C. S. The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology*, 1975, *31*, 674–685.

Dweck, C. S., & Bush, E. S. Sex differences in learned helplessness: I. Differential debilitation with peer and adult evaluators. *Developmental Psychology*, 1976, *12*, 147–156.

Dweck, C. S., & Reppucci, N. D. Learned helplessness and reinforcement responsibility in children. *Journal of Personality and Social Psychology*, 1973, *25*, 109–116.

Eastman, C. Behavioral formulations of depression. *Psychological Review*, 1976, *83*, 277–291.

Eisenberger, R., Park, D. C., & Frank, M. Learned industriousness and social reinforcement. *Journal of Personality and Social Psychology*, 1976, *33*, 227–232.

Eswara, H. S. Administration of reward and punishment in relation to ability, effort, and performance. *Journal of Social Psychology*, 1972, *87*, 139–140.

Feather, N. T. Attribution of responsibility and valence of success and failure in relation to initial confidence and task performance. *Journal of Personality and Social Psychology*, 1969, *13*, 129–144.

Feather, N. T., & Simon, J. G. Attribution of responsibility and valence of outcome in relation to initial confidence and success and failure of self and other. *Journal of Personality and Social Psychology*, 1971, *18*, 173–188. (a)

Feather, N. T., & Simon, J. G. Causal attributions for success and failure in relation to expectations of success based on selective or manipulated control. *Journal of Personality*, 1971, *39*, 522–541. (b)

Fontaine, G. Social comparison and some determinants of expected performance in a novel situation. *Journal of Personality and Social Psychology*, 1974, *29*, 487–496.

Fosco, E., & Geer, J. Effects of gaining control over aversive stimuli after differing amounts of no control. *Psychological Reports*, 1971, *29*, 1153–1154.

Gatchel, R. J., Paulus, P. B., & Maples, C. W. Learned helplessness and self-reported affect. *Journal of Abnormal Psychology*, 1975, *84*, 732–734.

Gatchel, R. J., & Proctor, J. D. Physiological correlates of learned helplessness in man. *Journal of Abnormal Psychology*, 1976, *85*, 27–34.

Geer, J., Davison, G. C., & Gatchel, R. J. Reduction of stress in humans through nonveridical perceived control of aversive stimulation. *Journal of Personality and Social Psychology*, 1970, *16*, 731–738.

Gilmor, T. M., & Minton, H. L. Internal versus external attribution of task performance as a function of locus of control, initial confidence, and success–failure outcome. *Journal of Personality*, 1974, *42*, 159–174.

Glass, D. C., & Singer, J. E. *Urban stress experiments on noise and social stressors*. New York: Academic Press, 1972.

Griffith, M. Effects of noncontingent success and failure on mood and performance. *Journal of Personality*, 1977, *45*, 442–457.

Hammen, C. L., & Krantz, S. Effect of success and failure on depressive cognitions. *Journal of Abnormal Psychology*, 1976, *85*, 577–586.

Hiroto, D. S. Locus of control and learned helplessness. *Journal of Experimental Psychology*, 1974, *102*, 187–193.

Hiroto, D. S., & Seligman, M. E. P. Generality of learned helplessness in man. *Journal of Personality and Social Psychology*, 1975, *31*, 311–327.

Jenkins, H. M., & Ward, W. C. Judgement of contingency between responses and outcomes. *Psychological Monographs*, 1965, *79*(1, Whole No. 594).

Kiesler, D. J. Experimental designs in psychotherapy research. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.

Klein, D. C., Fencil-Morse, E., & Seligman, M. E. P. Learned helplessness, depression, and the attribution of failure. *Journal of Personality and Social Psychology*, 1976, *33*, 508–516.

Klein, D. C., & Seligman, M. E. P. Reversal of performance deficits and perceptual deficits in learned helplessness and depression. *Journal of Abnormal Psychology*, 1976, *85*, 11–26.

Krantz, D. S., Glass, D. C., & Snyder, M. L. Helplessness, stress level, and the coronary-prone behavior pattern. *Journal of Experimental Social Psychology*, 1974, *10*, 284–300.

Levine, M. Hypothesis theory and nonlearning despite ideal S–R reinforcement contingencies. *Psychological Review*, 1971, *78*, 130–140.

Levis, D. J. Learned helplessness: A reply and an alternative S–R interpretation. *Journal of Experimental Psychology: General*, 1976, *105*, 47–65.

Loeb, A., Beck, A. T., & Diggory, J. Differential effects of success and failure on depressed and nondepressed patients. *Journal of Nervous and Mental Disease*, 1971, *152*, 106–114.

Maier, S. F., & Seligman, M. E. P. Learned helplessness: Theory and evidence. *Journal of Experimental Psychology*, 1976, *105*, 3–46.

Maier, S. F., Seligman, M. E. P., & Soloman, R. L. Pavlovian fear conditioning and learned helpless-

ness. In B. A. Campbell & R. M. Church (Eds.), *Punishment.* New York: Appleton-Century-Crofts, 1969.

Maier, S. F., & Testa, T. Failure to learn to escape by rats previously exposed to inescapable shock is partly produced by associative interference. *Journal of Comparative and Physiological Psychology,* 1975, *88,* 554–564.

Malmo, R. B. Finger sweat prints in differentiation of low and high incentive. *Psychophysiology,* 1965, *1,* 231–240.

McCarron, L. T. Psychophysiological discriminants of reactive depression. *Psychophysiology,* 1973, *10,* 223–230.

McMahan, I. D. Relationship between causal attributions and expectancy of success. *Journal of Personality and Social Psychology,* 1973, *28,* 108–114.

Miller, W. R., & Seligman, M. E. P. Depression and the perception of reinforcement. *Journal of Abnormal Psychology,* 1973, *82,* 62–73.

Miller, W. R., & Seligman, M. E. P. Depression and learned helplessness in man. *Journal of Abnormal Psychology,* 1975, *84,* 228–238.

Miller, W. R., & Seligman, M. E. P. Learned helplessness, depression, and the perception of reinforcement. *Behaviour Research and Therapy,* 1976, *14,* 7–17.

Miller, W. R., Seligman, M. E. P., & Kurlander, H. M. Learned helplessness, depression, and anxiety. *Journal of Nervous and Mental Disease,* 1975, *161,* 347–357.

Moos, R. H. Sources of variance in responses to questionnaires and in behavior. *Journal of Abnormal Psychology,* 1969, *74,* 405–412.

Mould, D. E. Differentiation between depression and anxiety: A new scale. *Journal of Consulting and Clinical Psychology,* 1975, *43,* 592.

Overmier, J. B. Interference with avoidance behavior: Failure to avoid traumatic shock. *Journal of Experimental Psychology,* 1968, *78,* 340–343.

Overmier, J. B., & Seligman, M. E. P. Effects of inescapable shock upon subsequent escape and avoidance learning. *Journal of Comparative and Physiological Psychology,* 1967, *63,* 23–33.

Padilla, A. M., Padilla, C., Ketterer, T., & Giacolone, D. Inescapable shocks and subsequent escape/avoidance conditioning in goldfish, *Carassius auratus. Psychonomic Science,* 1970, *20,* 295–296.

Parducci, A. Range-frequency compromise in judgment. *Psychological Monographs,* 1963, *77*(2, Whole No. 565).

Parducci, A. Category judgment: A range-frequency model. *Psychological Review,* 1965, *72,* 407–418.

Parducci, A. The relativism of absolute judgments. *Scientific American,* 1968, *219,* 84–90.

Phares, E. J. Expectancy change in chance and skill situations. *Journal of Abnormal and Social Psychology,* 1957, *54,* 339–342.

Rest, S., Nierenberg, R., Weiner, B., & Heckhausen, H. Further evidence concerning the effects of perceptions of effort and ability on achievement evaluation. *Journal of Personality and Social Psychology,* 1973, *28,* 187–191.

Roth, S., & Bootzin, R. R. The effects of experimentally induced expectancies of external control: An investigation of learned helplessness. *Journal of Personality and Social Psychology,* 1974, *29,* 253–264.

Roth, S., & Kubal, L. Effects of noncontingent reinforcement on tasks of differing importance: Facilitation and learned helplessness. *Journal of Personality and Social Psychology,* 1975, *32,* 680–691.

Rotter, J. B. *Social learning and clinical psychology.* Englewood Cliffs, N. J.: Prentice-Hall, 1954.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs,* 1966, *80*(1, Whole No. 609).

Rotter, J. B., Chance, J., & Phares, E. J. (Eds.). *Applications of a social learning theory of personality.* New York: Holt, Rinehart & Winston, 1972.

Rotter, J. B., Liverant, S., & Crowne, D. P. The growth and extinction of expectancies in chance controlled and skilled tasks. *Journal of Psychology,* 1961, *52,* 161–177.

Seligman, M. E. P. Fall into helplessness. *Psychology Today,* June 1973, 43–48.

Seligman, M. E. P. Depression and learned helplessness. In R. J. Friedman & M. M. Katz (Eds.), *The psychology of depression: Contemporary theory and research.* New York: Wiley, 1974.

Seligman, M. E. P. *Helplessness: On depression, development, and death.* San Francisco: Freeman, 1975.

Seligman, M. E. P., & Groves, D. Non-transient learned helplessness. *Psychonomic Science,* 1970, *19,* 191–192.

Seligman, M. E. P., & Maier, S. F. Failure to escape traumatic shock. *Journal of Experimental Psychology,* 1967, *74,* 1–9.

Seligman, M. E. P., Maier, S. F., & Geer, J. H. The alleviation of learned helplessness in the dog. *Journal of Abnormal Psychology,* 1968, *73,* 256–262.

Sherrod, D. R., & Downs, R. Environmental determinants of altruism: The effects of stimulus overload and perceived control on helping. *Journal of Experimental Social Psychology,* 1974, *10,* 468–479.

Tennen, H., & Eller, S. J. Attributional components of learned helplessness and facilitation. *Journal of Personality and Social Psychology,* 1977, *35,* 265–271.

Thomas, E., & Balter, A. Learned helplessness: Amelioration of symptoms by cholinergic blockage at the septum. *Science,* in press.

Thornton, J. W., & Jacobs, P. D. Learned helplessness in human subjects. *Journal of Experimental Psychology,* 1971, *87,* 369–372.

Thornton, J. W., & Jacobs, P. D. The facilitating effects of prior inescapable unavoidable stress on intellectual performance. *Psychonomic Science,* 1972, *26,* 185–187.

Underwood, B. J. Individual differences as a crucible in theory construction. *American Psychologist*, 1975, *30*, 128–134.

Velten, E. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 1968, *6*, 473–482.

Weiner, B. (Ed.). *Achievement motivation and attribution theory*. Morristown, N.J.: General Learning Press, 1974.

Weiner, B., Frieze, I., Kukla, A., Reed, L., Rest, S., & Rosenbaum, R. M. *Perceiving the causes of success and failure*. Morristown, N.J.: General Learning Press, 1971.

Weiner, B., & Kukla, A. An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology*, 1970, *15*, 1–20.

Weiner, B., Nierenberg, R., & Goldstein, M. Social learning (locus of control) versus attributional (causal stability) interpretations of expectancy of success. *Journal of Personality*, 1976, *44*, 52–68.

Wener, A. E., & Rehm, L. P. Depressive affect: A test of behavioral hypotheses. *Journal of Abnormal Psychology*, 1975, *84*, 221–227.

Wortman, C. B., & Brehm, J. W. Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 8). New York: Academic Press, 1975.

Wortman, C. B., Panciera, L., Shusterman, L., & Hibscher, J. Attributions of causality and reactions to uncontrollable outcomes. *Journal of Experimental Social Psychology*, 1976, *12*, 301–316.

Zuckerman, M., Lubin, B., & Robins, S. Validation of the Multiple Affect Adjective Check List in clinical situations. *Journal of Consulting Psychology*, 1965, *29*, 594.

Received September 29, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 119–138

The Concepts of Alienation and Involvement Revisited

Rabindra N. Kanungo
Faculty of Management, McGill University, Montreal, Canada

In an attempt to clarify and operationalize the concepts of alienation and involvement, this article critically examines several current sociological and psychological approaches to the concepts. Several common sources of confusion surrounding the treatment of the concepts are identified. A motivational formulation of the concepts is suggested to achieve greater parsimony and integration of the diverse sociological and psychological thinking on the issue. Finally, some major implications of this motivational formulation for future empirical studies on alienation and involvement at work and in other situations are indicated.

In the social science literature of the past two decades, one encounters very often the usage of the concept *alienation* and its obverse, *involvement* (Johnson, 1973). These concepts have been used by sociologists, psychologists, political scientists, theologians, philosophers, and historians to describe and explain various contemporary social phenomena. In fact, the terms alienation and involvement have been used so often and in so many contexts that they have acquired an aura of equivocality. As Seeman (1971) pointed out, the concept of alienation has been "popularly adopted as the signature of the present epoch. It has become routine to define our troubles in the language of alienation and to seek solutions in those terms. But signatures are sometimes hard to read, sometimes spurious, and sometimes too casually and promiscuously used. They ought to be examined with care" (p. 135). Similar concern was expressed by Johnson (1973), who characterized the concept of alienation as being capable of carrying a great deal of feeling "in an inexplicit, perplexing and deeply annoying way" (p. 28). Although in recent years many psychologists and sociologists have attempted to demystify and to operationalize the concept (Lawler & Hall, 1970; Lodahl & Kejner, 1965; Saleh & Hosek, 1976; Seeman, 1971; Vroom, 1962), none of them seem to offer a scientifically organized and meaningful view of the concept that could have broad generality across cultures. Although most of the researchers have tried to explain the phenomena of alienation and involvement in social psychological terms (Clark, 1959; Lawler & Hall, 1970; Seeman, 1959), the language they have used seems to have created more confusion than clarity. Sociological and psychological explanations of the phenomena seem to run parallel courses of their own without any serious attempt at integration. In fact, if one puts together the various explanations of the phenomena advanced by these writers, one ends up with greater conceptual fuzziness rather than clarity or understanding. If we take seriously Seeman's (1971) call for a careful examination of the concept for better clarity and rigor, we ought to seek a reformulation of the issue. The present article is an effort in this direction.

First, the article traces the manifold nature of the concept of alienation as it has been viewed by several researchers in this area. Second, the article identifies the major sources of confusion surrounding this concept. And finally, the article presents a motivational formulation of alienation and a comprehen-

The author wishes to thank George Strauss and Thomas Reynolds of the University of California, Berkeley and Gerald Gorn of McGill University for their helpful comments on an earlier draft of the article.

Requests for reprints should be sent to Rabindra N. Kanungo, Faculty of Management, McGill University, Samuel Bronfman Building, 1001 Sherbrook Street West, Montreal, Quebec, Canada H3A 1G5.

sive conceptual framework that could be used in future research. Such a framework is aimed at integrating the sociological and the psychological thinking on the issue and at providing a more complete understanding of the complex phenomena of alienation and involvement in a parsimonious way. It must be pointed out that the terms *alienation* and *involvement* are used here to indicate bipolar states of the same phenomena. Furthermore, since the phenomena have been studied empirically mostly in work situations, I have chosen to discuss specifically the nature of alienation and involvement at work in some detail. This choice however does not imply that one should limit the use of the concepts of alienation or involvement to work situations alone. The framework suggested in this article is in fact intended to be used by researchers not only to study work alienation phenomena but also to study alienation in family life, in religious contexts, and for that matter in any other specific aspects of one's environment.

Alienation and Involvement: Some Earlier Formulations

Although contemporary social scientists often consider the state of alienation in individuals as distinctly a postindustrial phenomenon, theologians and philosophers claim otherwise. As Johnson (1973) pointed out, social alienation as an observed phenomenon is quite ancient and the term alienation is an antique one. Theologians take the credit for using the term as an explanatory concept. Alienation, according to them, refers to states of separation of human beings from God, from their own bodies, from their fellow human beings, and from their institutions. This interpretation is clearly reflected in the use of dualism of body and soul in theological writings. Essentially, the theologians observed and explained the meaninglessness of human existence in terms of spiritual alienation or separation from God and moral principles. Although alienation as a psychological state of the individual (or as a collective social phenomenon) has been recognized for centuries, the scientific treatment of the concept with regard to its nature and its effect was attempted first by empirically oriented sociologists and then, more recently, by social psychologists. Thus the concept has lived through two distinct traditions, the rational and the empirical. The rational tradition of the concept comes largely from the writings of theologians (Macquarrie, 1973) and philosophers (Denise, 1973). The empirical tradition results from the recent works of sociologists and psychologists. The main focus of the present article is a critical review of the empirical tradition.

In dealing with states of alienation in the spiritual life of individuals, theological approaches emphasized the idea that there can be alienation of different sorts depending on what elements of one's environment one is separated from (such as God, one's own body, other people, etc.). Following a similar line of reasoning, the social scientists of today talk of different kinds of alienation such as job alienation, organizational alienation, urban alienation, family alienation, and so on. From the point of view of conducting empirical research, social scientists consider the study of alienation in relation to single, well-defined environmental elements (such as job, family, etc.) to be more fruitful than the study of alienation in a global sense (Clark, 1959; Seeman, 1971). The theological and philosophical approaches to the concept influenced the thinking of modern social scientists in yet another way. The more recent writers have identified the core meaning of the concept of alienation as a dissociative state of the individual (a cognitive sense of separation) in relation to some other element in his or her environment (Schacht, 1970). In the following section, an overview of the sociological and the psychological approaches to the concept is presented.

The Sociological Approach

The contributions of sociologists in explaining the nature of alienation have been the most extensive. In the classics of sociology, such as in the writings of Marx, Weber, and Durkheim, the concept of alienation received very comprehensive treatment. Al-

though Rousseau was the first to provide sociological treatment of the concept, it was Marx and later Weber who put the concept on firmer analytic ground. The evidence of their powerful influence still persists in the works of contemporary sociologists (Seeman, 1959, 1971).

If theologians identified the basic psychological state of alienation in men in their spiritual lives, Marx identified it in their material, working lives. According to Marx (1844/1932), labor or working on a job is the "existential activity of man, his free conscious activity—not a means of maintaining his life but for developing his universal nature" (pp. 87–88). Thus, ideally a state of work involvement can result when the work situation elicits job behavior that is perceived to be (a) voluntary, (b) not instrumental in satisfying basic physical needs, (c) instrumental in satisfying Maslow-type (1954) higher order needs such as the need for self-realization or self-actualization, and (d) conducive to developing individuals' abilities to their fullest potential. In the absence of such perceptions, the individual worker is bound to experience a state of alienation from work. Most work setups according to Marx provide conditions that alienate workers rather than involve them. Marx identified two major job conditions that are responsible for alienation among workers. They are (a) separation of workers from the products of their labor and (b) separation of workers from the means of production. The first job condition implies that the product is perceived as not belonging to the worker. The worker also perceives that he or she cannot influence the disposition or quality of the product. Thus he or she lacks a sense of ownership and of control over the product and its quality. The second job condition implies that the worker perceives a lack of control over the function of the machines and other means of production. Finding that he or she has no control over working life, the worker is bound to be estranged or to separate working life from the rest of his or her existence over which Marx assumes the worker has complete control.

From the preceding discussion, it becomes obvious that it is the lack of *autonomy* and *control* of one's behavior and its effects that defines the Marxian concept of alienation. If one translates the alienation state of the worker into motivational terms, it becomes quite clear that Marx intended to measure alienation in terms of the satisfaction of a single set of needs, the ego needs for independence, achievement, and power. In the Marxian formulation, the role of other human needs, such as the physical and the social ones, has been completely disregarded, as if the needs do not constitute a part of one's self or perhaps constitute a very insignificant part exerting almost no influence in causing states of alienation. This interpretation of the Marxian formulation may appear oversimplified, but it is clearly reflected in the following quotation from Marx (1844/1932):

> What constitutes alienation of labor? First, that work is *external* to the worker, that it is not part of his nature; and that, consequently, he does not fulfill himself in his work but denies himself, has a feeling of misery rather than well-being, does not develop freely his mental and physical energies but is physically exhausted and mentally debased. The worker therefore feels himself at home only during his leisure time, whereas at work he feels homeless. His work is not voluntary but imposed, forced labor. It is not the satisfaction of a need, but only a means for satisfying other needs. (pp. 85–86)

One may notice the assumptions Marx makes while defining the state of alienation. Clearly he emphasizes the worker's experience of frustration of his or her autonomy and control needs at work, and whenever these needs are frustrated, Marx considers work to be external to the worker's self.

Another assumption made by Marx in the quotation relates to the instrumental and consummatory properties of job behavior. According to Marx, job behavior can be either instrumental activity that satisfies basic physical human needs or it can be consummatory activity. In the former sense job behavior is viewed as means to an end (satisfaction of extrinsic needs), and in the latter sense it is viewed as an end in itself. Theories of human motivation suggest that human behavior is purposive; it has directionality; it is initiated by need states; and it is always instrumental in satisfying these need states. An individual's job behavior is also purposive; it is aimed at

satisfying both the extrinsic and intrinsic need states (Lawler, 1973) of the individual. However, when Marx wrote of job behavior as an end in itself (reflecting a state of involvement), he did not recognize that such behavior is also instrumental in satisfying a set of intrinsic human needs.

Weber's treatment of the concept of alienation is very similar to that of Marx. As Gerth and Mills (1946) put it, "Marx's emphasis upon the wage worker as being 'separated' from the means of production becomes, in Weber's perspective, merely one special case of a universal trend. The modern soldier is equally 'separated' from the means of violence, the scientist from the means of enquiry, and the civil servant from the means of administration" (p. 50). Weber's exposure to the American way of life (political democracy and economic capitalism) and his study of the Protestant religion convinced him that the spirit of the Protestant work ethic is the key to the realization of man's potentialities to the fullest extent. As Gerth and Mills wrote, Weber was impressed by the "grandiose efficiency of a type of man, bred by free associations in which the individual had to prove himself before his equals, where no authoritative commands, but autonomous decisions, good sense, and responsible conduct train for citizenship" (p. 18). Such is the image Weber had of an involved worker. Like Marx, Weber also placed emphasis on the freedom to make one's own decisions, on assuming personal responsibility, and on proving one's worth through achievement at work. Translated into motivational terms, this implies that if the work setup cannot provide an environment that satisfies the needs for individual autonomy, responsibility, and achievement, it will create a state of alienation in the worker.

Unlike Marx and Weber, who viewed alienation as resulting primarily from perceived lack of freedom and control at work, Emile Durkheim, the French sociologist, saw it as a consequence of a condition of *anomie*, or the perceived lack of socially approved means and norms to guide one's behavior for the purpose of achieving culturally prescribed goals (Blauner, 1964; Durkheim, 1893; Shepard, 1971). The condition of anomie is often considered a postindustrial phenomenon. As Blauner (1964) observed, industrialization and urbanization of modern society have "destroyed the normative structure of a more traditional society and up-rooted people from the local groups and institutions which had provided stability and security" (p. 24). No longer being able to feel a sense of security and belonging, modern men and women find themselves isolated from others. This form of social alienation often results in normlessness and in its collective form manifests itself in various forms of urban unrest. In social psychological terms, this variant of alienation seems to stem from the frustration of social and security needs, the need to belong to groups for social approval and social comparison (Festinger, 1954; Maslow, 1954). The social psychological processes that explain how this form of alienation comes about are discussed later in the article.

The strong impact of Marx, Weber, and Durkheim is quite evident in contemporary sociological writings on the subject of alienation and involvement. For instance, Dubin (1956) defined involvement as *central life interest*. According to him, a job-involved person is one who considers work to be the most important part of his or her life and engages in it as an end in itself. A job-alienated person, on the other hand, engages in work in a purely instrumental fashion and perceives work as providing financial resources for more important off-the-job activities. Faunce (Note 1) also considered job involvement as a commitment to a job in which successful performance is regarded as an end in itself rather than as a means to some other end. For both Dubin and Faunce, the concepts of involvement and alienation are intimately related to the Protestant work ethic, the moral value of work, and personal responsibility as conceived by Weber.

In an attempt to clarify the concept of alienation, Seeman (1959, 1971) has proposed five different variants of the concept: powerlessness, meaninglessness, normlessness, isolation, and self-estrangement. According to Seeman, each variant refers to a different subjectively felt psychological state of the individual, caused by different environmental con-

ditions. Several other researchers, particularly Blauner (1964) and Shepard (1971), have used Seeman's classification and have tried to provide operational measures of the different categories of alienation at work. They also have suggested the antecedent physical and social conditions that produce each state of alienation.

Alienation in the form of *powerlessness* in the most general sense refers to a perceived lack of control over important events that affect one's life. Seeman (1959) used this variant of alienation to explain and describe men and women's alienation from the larger social order. An individual's inability to control and influence political systems, industrial economies, or international affairs may create a sense of powerlessness in him or her. Alienation in the sense of powerlessness also has been observed in job situations. For instance, Shepard (1971) described powerlessness at work as "the perceived lack of freedom and control on the job" (pp. 13–14). Blauner (1964) expressed similar views when he stated that "the non-alienated pole of the powerlessness dimension is freedom and control" (p. 16). According to Blauner, the powerlessness variant of alienation at work results from the mechanization process that controls the pace of work and thus limits workers' free movements. If one analyzes the sociological concept of powerlessness in motivational terms, it becomes obvious that if a situation constantly frustrates an individual's needs for autonomy and control, it will create in him or her a state of alienation of this type.

The second type of alienation is identified as a cognitive state of *meaninglessness* in the individual. In such a state, the individual is unable to predict social situations and the outcomes of his or her own and others' behavior. In the work setup such a state results from increasing specialization and division of labor. When the work process is broken down into simple minuscule tasks, and when such simple tasks involve no real responsibility and decision making, the work situation robs the worker of any sense of purpose. The job becomes meaningless for the worker. Meaninglessness of work may also result when the worker is not able to see the relation of his or her work to the total system of goals of the organization (Blauner, 1964; Shepard, 1971). Translated into motivational terms, this implies that continued frustration of an individual's needs for assuming personal responsibility and for gaining greater competence on the job (by being more knowledgeable about the environment for the sake of influencing it) causes this type of alienation. It may be noted that both the powerlessness and the meaninglessness interpretations of work alienation bear the mark of the Marxian belief that lack of control and freedom over the work process is the main cause of alienation.

The two other forms of alienation suggested by Seeman (1959) have their roots in Durkheim's (1893) description of anomie. Anomie refers to the perceived conditions of one's social environment, such as the perception of a breakdown of social norms regulating individual conduct in modern societies. The two forms of alienation that result from such perceived conditions of one's social environment are *normlessness* and *isolation.* An individual may develop a sense of normlessness when he or she finds that previously approved social norms are no longer effective in guiding behavior for the attainment of personal goals. In other words, the individual finds that to achieve given goals it is necessary to use socially unapproved behavior. Finding that he or she can no longer share the normative system because of its ineffectiveness, the person may develop norms of his or her own to guide behavior. Because his or her norms are different from those of others, the individual may eventually perceive himself or herself as being separate from society and its normative system. The dissociation of oneself from others results in the perception of social isolation. The dissociation of oneself from social norms results in normlessness or cultural estrangement. Alienation in the sense of social isolation and cultural estrangement refers to the perceived states of loneliness and rootlessness respectively (Seeman, 1971). It may be noticed that these two variants of alienation are related because they stem from the same basic condition of anomie.

States of loneliness and rootlessness have also been identified in work environments.

Blauner (1964), for instance, suggested that these forms of social alienation may be manifested on the job due to the lack of social integration of the worker. When an organization does not provide the worker any opportunity for developing a sense of membership or belonging in the social system, the worker is bound to show a sense of isolation from the system and its goals. From a motivational point of view, the two variants of social alienation, isolation and normlessness, seem to be based on two different social needs of the individual. Continuous frustration of the membership or belonging need of the individual may be the crucial determinant of the isolation form of alienation. The normlessness form of alienation, however, is determined by continuous frustration of another social need, the need to evaluate oneself through social comparison (Festinger, 1954). In the context of social influence theories, social psychologists (Jones & Gerard, 1967) have postulated two major kinds of influences that groups exert on the individual. They are referred to as the normative and the informational social influences. By being a member of the group and by adhering to the group norms, the individual fulfills his or her need to belong, to love, and to be loved by others. When, however, the group norms are perceived to be too restrictive and in conflict with the individual's personal goals, they cease to influence the individual. The group loses its normative influence on the individual. The person becomes an isolate in relation to the group. He or she perceives himself or herself as one who no longer belongs to the group and no longer is loved by others in the group. Such a psychological state can be identified as the isolation form of alienation. The individual also depends on the group norms for self-evaluation and for evaluating his or her abilities and opinions (Festinger, 1954). Group norms generally provide the person with information on how to behave, on what is right and what is wrong. When the individual finds that group norms do not provide useful information for self-evaluation, he or she may separate himself or herself from these norms and experience a state of normlessness. Thus, in terms of social influence theory, the two variants of social alienation result from the failure of the groups to exercise the two forms of social influence, normative and informational.

The final variant of alienation proposed by sociologists is *self-estrangement*. In many ways the characterization of this category of alienation has posed problems for sociological thinkers. Seeman (1971) admits that it is an "elusive idea" (p. 136), but then goes on to operationalize it. According to Seeman, a person is self-estranged when he or she is engaged in activity that is not rewarding in itself, but is instrumental (a means to an end) in satisfying extrinsic needs such as the needs for money, security, and so on. Following Seeman (1959), Shepard (1971) considers instrumental work orientation, or the degree to which one works for extrinsic-need satisfaction, to be an index of the self-estrangement kind of alienation in the work setup. Blauner (1964) suggests that a job encourages self-estrangement if it does not provide opportunity for expressing "unique abilities, potentialities, or personality of the worker" (p. 26). In motivational terms, Blauner's observation means that whenever the individual finds his or her environment (job or otherwise) lacking in opportunities for the satisfaction of self-actualization needs (Maslow, 1954) through expression of his or her potentialities, he or she experiences a state of self-estrangement. Following Marx, many contemporary sociologists believe that self-estrangement is the heart of the alienation concept, as if all other forms of alienation eventually result in self-estrangement. Blauner (1964) attests to this belief in the following remark: "When work activity does not permit control [powerlessness], evoke a sense of purpose [meaninglessness], or encourage larger identification [isolation], employment becomes simply a means to the end of making a living" (p. 3).

Characteristics of the Sociological Approach

At this point it may be helpful to identify some dominant considerations that have guided most sociological treatments of the concept of alienation.

First, one notices a stronger emphasis in sociological writings on the analysis and

measurement of the state of alienation than on the analysis and measurement of the state of involvement. In a sense, sociologists have focused their attention on the negative side of the issue with a clinical perspective on social systems. Thus, they have been more concerned with the diagnosis of states of alienation in social systems and consequent social maladies than with the identification of conditions for social involvement and growth. Like Freudian psychologists who attempt to explain human nature through an analysis of pathological psychological states, sociologists, taking the lead from Marx, have emphasized the analysis of alienation and resulting pathological states to explain the nature of social systems. In the same way as the Freudian influence in psychology delayed the formulation of growth theories of personality and motivation (Maslow, 1954; Allport, 1961), the Marxian influence in sociology may have retarded the progress of sociological theories in understanding better the nature of healthy and growing social systems. As is discussed later, unlike the sociological approaches outlined above, the current psychological approaches to the issue are trying to attack the problem from the positive side through the study of the concept of involvement.

The second consideration that has dominated various sociological treatments of alienation is their emphasis on studying alienation in groups and social systems. The level of analysis of the concept in most sociological approaches has been at the social system level rather than at the individual level. This has created measurement problems. Although sociologists often talk of the frequency of volatile activism, of suicide rates, of crime rates, and so on as indices of alienation in social systems, they find it hard to establish and theoretically justify the validity and the reliability of these measures. The records on such social maladies are notoriously unreliable. Very often incidents of activism, crime, and suicide go unreported. Even if the incidents are recorded accurately, it is often difficult to infer from these data states of alienation in individual persons. For instance, an activist in his or her desire to bring about changes in the social system may be showing signs of greater involvement in the social system than would an apathetic conformist.

Third, sociological approaches generally describe the state of alienation not in specific behavioral terms, but in terms of epiphenomenal categories. As Johnson (1973) pointed out, alienation is seen as *"an epiphenomenal abstraction,* collectively summarizing a series of specific behaviors and categorizing them as 'loneliness,' 'normlessness,' 'isolation,' etc." (p. 40). Such epiphenomenal descriptions of the concept may have the flavor of intellectual romanticism, but have very little scientific value because they pose problems of empirical verification. The concept of alienation as an epiphenomenal abstraction tends to carry excess meaning and therefore eludes precise measurement. Besides, such an abstraction merely describes alienation; it does not explain it.

Finally, most sociological approaches consider the presence of individual autonomy, control, and power over the environment as basic preconditions for removing the state of alienation.

The Psychological Approach

The psychological approach to the concept of alienation has been somewhat sketchy compared to the sociological approach described earlier. In psychological literature, the treatment of the concept does not have as long and as rich a tradition as in sociology. The interest in the concept is very recent among psychologists, and they have essentially taken an empirical (and exploratory) approach to the study of the problem. Development of psychological theories to explain the phenomenon of alienation is simply absent from the literature. Furthermore, in contrast to the sociological approach, psychologists have attempted to analyze the nature of alienation only in the limited context of job situations. Unlike sociologists, psychologists have studied the problem of alienation from the point of view of job involvement and have attempted to define and measure involvement at work rather than alienation at work.

In trying to explain the nature of job involvement, psychologists have concentrated

on the analysis of specific motivational states of the individual in work situations. Psychological explanations are based on motivation theories and therefore tend to emphasize the need-satisfying qualities of the job as basic determinants of job involvement. For instance, Vroom (1962) proposed that a person's attempts to satisfy his or her needs for self-esteem through work on the job lead to job involvement. In his study, "The degree of job involvement for a particular person was measured by his choice of 'ego' rather than extrinsic factors in describing the sources of satisfaction and dissatisfaction on the job" (p. 161). Vroom seems to emphasize intrinsic-need satisfaction as the essential condition for higher job involvement. In his view, higher autonomy extended to the individual results in higher ego involvement, which in turn leads to a higher level of job performance.

Lodahl and Kejner (1965) proposed two definitions of job involvement. One of their definitions states that "job involvement is the degree to which a person is identified psychologically with his work, or the importance of work in his total self-image" (p. 24). Such a psychological state of identification with work may result partly from early socialization training during which the individual may internalize the value of the goodness of work. Lodahl and Kejner (1965) recognized this possibility. They stated that the concept of job involvement "operationalizes the 'protestant ethic' and because it is a result of the introjection of certain values about work into the self, it is probably resistant to changes in the person due to the nature of a particular job" (p. 25). Lodahl and Kejner also provided another definition of job involvement; this definition states that job involvement is "the degree to which a person's work performance affects his self-esteem" (p. 25). These two definitions are quite distinct, and Lodahl and Kejner made no attempt in their study to show how the two are related. In fact, the questionnaire measure of job involvement they developed includes items reflecting both definitions. Use of their questionnaire measure in job involvement research therefore provides data that are hard to interpret.

Recently Rabinowitz and Hall (1977) critically reviewed the work of several researchers who have made use of the definitions of job involvement mentioned above. Their review clearly suggests that there is a great deal of confusion and ambiguity in theories about job involvement. Furthermore, as these authors pointed out, "The confusion does not stop at the theoretical level, but rather continues in the empirical studies of involvement" (p. 267).

Weissenberg and Gruenfeld (1968) investigated the relationship between satisfaction with various job factors and job involvement. They concluded that increased job involvement is positively related to satisfaction with motivators or job-content factors (Herzberg, 1966) such as achievement, responsibility, independence, and so forth. These motivators tend to satisfy the intrinsic needs of the individual. The extrinsic needs, however, are satisfied through job-context factors such as company policies, nature of supervision, salary, benefits, and working conditions. According to these researchers, satisfaction with the job-context factors is unrelated to job involvement, but the latter can be predicted from the satisfaction with the motivators in the job.

Lawler and Hall (1970) for the first time distinguished the psychological state of job involvement from two other psychological states of the worker. According to Lawler and Hall, job involvement is different from both intrinsic motivation on the job and job satisfaction. Intrinsic motivation refers to a state of the individual in which satisfaction of the intrinsic needs is contingent upon appropriate job behavior and in which job satisfaction results from satisfaction of the needs of the individual through the attainment of job outcomes without any regard to the contingencies of the outcomes. Lawler and Hall argue in favor of the definition of job involvement suggested by Lodahl and Kejner (1965): Job involvement is seen in terms of psychological identification with work or the importance of work to one's total self-image. In general Lawler and Hall suggest that job involvement refers to the "degree to which a person's total work situation is an important part of his life. The job-involved person is

one who is affected very much personally by his whole job situation, presumably because he perceives his job as an important part of his self-concept and perhaps as a place to satisfy his important needs (e.g., his need for self-esteem)" (pp. 310–311). It appears that in defining the concept of involvement, Lawler and Hall assumed that intrinsic or growth needs (Alderfer, 1972) are central to the self-concept of the individual. To emphasize the centrality of intrinsic needs, they pointed out that "the more the job is seen to allow the holder to influence what goes on, to be creative, and to use his skills and abilities, the more involved he will be in the job" (p. 310). In the same article, Lawler and Hall (1970) reiterated their position with the following remark: "Other things being equal, more people will become involved in a job that allows them control and a chance to use their abilities than will become involved in jobs that are lacking these characteristics" (p. 311).

Patchen (1970) identified three general conditions for job involvement. According to him, "Where people are highly motivated, where they feel a sense of solidarity with the enterprise, and where they get a sense of pride for their work, we may speak of them as highly 'involved' in their job" (p. 7). When Patchen talks of workers being highly motivated, he refers to their high levels of achievement need or to their wish to accomplish worthwhile things on the job. When he talks of workers' solidarity with the enterprise, he refers to their need for belonging to the organization. Finally, when he talks of workers' sense of pride, he refers to workers' feeling of high self-esteem. Thus, in Patchen's view, when a job provides opportunities for the satisfaction of one's achievement needs, belonging needs, and self-esteem needs, one experiences a greater degree of job involvement.

In a recent review of the psychological literature on job involvement, Rabinowitz and Hall (1977, p. 284) stressed that among other things, a job-involved person believes strongly in the Protestant ethic, has strong growth needs, and has a stimulating job that gives him or her a high degree of autonomy and an opportunity for participation.

In another review of the psychological literature on job involvement, Saleh and Hosek (1976) identified four different interpretations of the concept of involvement: "A person is involved (1) when work to him is a central life interest; (2) when he actively participates in his job; (3) when he perceives performance as central to his self-esteem; (4) when he perceives performance as consistent with his self-concept" (p. 215). The first interpretation of the concept of involvement in terms of central life interest (Dubin, 1956) is very similar to the interpretation offered by Lawler and Hall (1970). The main idea underlying this interpretation is that the psychological state of involvement with respect to an environmental entity (such as job, family, etc.) is a cognitive or perceived state of identification with that entity. The second interpretation of involvement in terms of participation suggests that the psychological state of involvement be viewed as behavioral acts of the individual directed toward the satisfaction of his or her needs for autonomy and control. Bass (1965), for instance, considered participative job behaviors such as making important job decisions, setting one's own work pace, and so on to be important indices of greater work involvement. The remaining two interpretations of involvement, namely, providing a sense of personal worth (Siegel, 1969) and reinforcing one's self-concept (Vroom, 1964), suggest that involvement may be viewed as the experience of satisfaction resulting from the fulfillment of the individual's self-esteem and self-actualization needs. From the results of their own factor analytic work, Saleh and Hosek (1976) concluded that job involvement is "the degree to which the person identifies with the job, actively participates in it, and considers his performance important to his self-worth. It is, therefore, a complex concept based on cognition, action, and feeling" (p. 223). It is interesting to note that to achieve conceptual clarity Lawler and Hall tried to differentiate the state of involvement from intrinsic motivation and job satisfaction, whereas Saleh and Hosek brought them all together again.

At this point it must be noted that there is one common thread that runs through all the

psychological formulations outlined above. All of them seem to emphasize that situations lacking in opportunity for the satisfaction of intrinsic needs of the individual such as self-esteem, achievement, autonomy, control, self-expression, and self-actualization will decrease the individual's involvement in them. Even the recent studies on central life interest in work settings, on organizational identification, and on organizational commitment (Dubin, Champoux, & Porter, 1975; Hall & Schneider, 1972; Hall, Schneider, & Nygren, 1970) reflect this bias. It seems as if the lack of intrinsic-need satisfaction is the basic condition for increasing work alienation. In this regard, psychologists seem to have followed the sociological tradition of considering the lack of individual freedom, power, and control as necessary preconditions of the psychological state of alienation.

Sources of Confusion Surrounding the Concepts

Five major sources of confusion can be identified in the literature on alienation and involvement. Not only in the theoretical treatment of the concepts but also in the operationalization of the concepts in empirical studies, one notices the presence of these sources of confusion. They have contributed to the exasperating conceptual ambiguity prevailing in the area, and it goes without saying that any meaningful scientific treatment of the concepts should guard against them.

The most common source of confusion is the application of the concepts sometimes to specific individuals and sometimes to groups of individuals. Particularly in sociological writings one finds the use of the concept of alienation sometimes to describe the psychological state of the individual and at other times to describe pathological states of large collectivities such as groups, organizations, and other socio-political systems. As Johnson (1973) correctly pointed out, "There is a difference in meaning between these two applications that is not merely the difference between singular and plural categories. The phenomenology and the meaning connected with individual states of alienation are different both in quality and significance from those connected with the social, interactional, and collective applications of the term" (p. 35). For example, to say that a worker is alienated can mean two things. It can suggest an instance of *collective experience* of worker alienation as reflected in absenteeism, tardiness, goldbrick, sabotage, and so on that results from the prevailing social and physical conditions (mechanization, impersonal control through rules and regulations, etc.) within the organization, or it can suggest an individual worker's *personal view* of his or her work that does not meet his or her salient needs (unique to the individual) regardless of how other workers view the situation. From a methodological standpoint, it is advisable to approach the study of alienation at the personal rather than at the collective level of experience. Measurement and interpretation of the collective experience of alienation are often difficult and confusing.

A second source of confusion stems from the fact that the concept of alienation has been described and measured in two different ways. Sometimes the term alienation is used to imply objective social conditions directly observed by others and later attributed to individuals and groups. Blauner (1964), for instance, considered mechanization and division of labor to be the alienating conditions, and people working under these conditions were assumed to be experiencing alienation. At other times, alienation has been interpreted as a subjective psychological state of the individual not detectable to outsiders but felt by the individual. Such a difference in the usage of the term has obvious implications for the operationalization of the concept. States of alienation measured through identification of objective conditions may not parallel the subjective measures of the concept. Mechanization and division of labor in an organization may be viewed by external observers as necessarily contributing to a state of alienation of the worker (powerlessness), but the worker may not perceive the situation in the same way. In fact, it is quite conceivable that for some workers (mentally and physically handicapped, unskilled, uneducated, and many belonging to developing countries) mechanization and division of labor may increase job involvement.

A third source of confusion (related to the preceding one) results from a failure to maintain the conceptual distinction between the *antecedent conditions* of alienation and the *consequent states* of alienation. Here the confusion results from mistaking the cause as the effect. As Josephson and Josephson (1973) remarked, "Durkheim's notion of anomie or normlessness can be regarded as an important *cause* of alienation but should not be confused with alienation as a state of mind. . . . By the same token, alienation should not be confused with 'social disorganization,' since estrangement may be found in highly organized bureaucracies" (p. 166). In spite of such warnings, both the sociological and the psychological formulations neglect to maintain the distinction between alienating conditions and alienating states. In fact, most empirical researchers have attempted to measure the state of alienation through indices of alienating conditions instead of directly measuring it (as if the two were equivalent). For instance, Seeman (1959) considered normlessness to be the perception of a social situation in which rules and norms regulating behavior have broken down. Such perceptions may be the antecedent conditions of the alienated state, but they cannot be identified with the alienated state itself. Likewise, isolation, meaninglessness, and powerlessness may describe different conditions or causes of alienation, but should not be equated with it. Even when self-estrangement was measured by Blauner (1964), he used several indices of alienating conditions on the job, such as whether the job met the worker's achievement needs. Shepard (1971) also measured the different forms of alienation suggested by Seeman (1959) by measuring various job conditions such as whether the job provided opportunity for participation and control (powerlessness), how the job fit into the total operation of the organization (meaninglessness), and the like. Clearly these kinds of questions probe into the assumed conditions or causes of alienation rather than into the state of alienation itself. In the psychological literature similar confusion may be noted. For instance, Saleh and Hosek (1976) have proposed a measure of job involvement that contains three distinct categories of items. The first category measures directly the state of alienation (e.g., with the item "The most important things I do are involved with my job"). The second category seems to index the antecedent conditions or presumed causes of alienation (e.g., with the item "How much chance do you get to do things your own way?"). Finally, a third category measures workers' behaviors and experiences that often (but not necessarily) result from the alienated state (e.g., the item "I avoid taking on extra duties and responsibilities in my work"). Thus Saleh and Hosek combine indices of causal conditions and effects of alienating states into one single instrument. Such an instrument cannot provide meaningful data on the state of alienation of the worker. Needless to say, for both conceptual clarity and effective methodology in empirical studies, the state of alienation needs to be identified and measured separately from its causes as well as its effects.

A fourth source of confusion results from the description of the state of alienation as being both a *cognitive* as well as an *affective* state of the individual. Most researchers have found it difficult to strip the concept of alienation from its negative affect. Traditionally, alienation has been associated with negative emotional states such as anger, dissatisfaction, and unpleasantness, and involvement has been associated with positive emotional states such as satisfaction and pleasantness. Many measures of alienation or involvement therefore contain items reflecting levels of satisfaction or dissatisfaction (e.g., the item "The major satisfaction in my life comes from my job" in Lodahl & Kejner, 1965). Recent empirical studies (Lawler & Hall, 1970; Seeman, 1971) clearly suggest that work involvement and job satisfaction are not the same thing, although they may be related to one another. It may be more useful to conceptualize the states of involvement or alienation as cognitive or belief states of identity or dissociation (separateness) than as psychological states necessarily associated with feelings of satisfaction or dissatisfaction. A cognitive state of dissociation may or may not accompany positive or negative affect under certain conditions. A highly involved worker under some conditions may

feel a high level of satisfaction with his or her work and under other conditions may experience deep dissatisfaction. In the future, empirical work needs to be done to identify conditions under which involvement and alienation are related to positive, negative, and neutral affective states.

Finally, some ambiguity regarding the concept of alienation has resulted from the confusion of two kinds of causation, *contemporaneous* and *historical*. Sometimes the state of alienation in an individual has been viewed as the result of the past history of the individual. For instance, Lodahl and Kejner (1965) suggested that work involvement of an individual is determined by the early socialization process during which the individual internalizes the values of the goodness of work or the Protestant ethic. In this sense, alienation from or involvement with work becomes a more stable characteristic of the individual, which he or she carries with him or her from one situation to another. Sociologists have viewed the historical causation of alienation in a slightly different way. Following Marx, many sociologists have considered job experience to be central to an individual's life. According to them, the longstanding social arrangements of technology, division of labor, and capitalist property institutions have created the state of alienation from work (Blauner, 1964). Since work is central to one's life, alienation from work necessarily leads to alienation from all other aspects of life. As Seeman (1971) put it, "Perhaps the most important thesis concerns the centrality of work experience, the imputation being that alienation from work 'is the core of all alienation' and that the consequences of alienated labor color the life space of the individual in a profound and disturbing way" (p. 135). The state of alienation has also been conceived as being caused by contemporaneous events. For instance, Lawler and Hall (1970) consider the job-involved person to be one who is "affected very much by his whole job situation, presumably because he perceives his job as an important part of his self-concept and perhaps as a place to satisfy his important needs (e.g., his need for self-esteem)" (pp. 310–311). These authors therefore consider the worker's present perceptions of the need-satisfying potentialities of the job to be a major determinant of the state of involvement. From the above discussion, it is apparent that a state of alienation or involvement with regard to the specific aspects of one's environment (such as work, family, religion, etc.) may be jointly caused by two sets of events—one historical and the other contemporaneous. Through the socialization process (cumulative learning and experience of the past) the individual may develop a set of relatively stable beliefs and values regarding work, family, and so on, and the present experiences with them may either reinforce the beliefs and values or modify them.

A Motivational Framework for the Study of Alienation and Involvement

The following discussion is a description of a conceptual approach that can be used to study the phenomena of alienation and involvement in any specific aspect of one's life. To provide an example, however, I have chosen to discuss work involvement or the work setup as the specific environmental entity toward which a state of alienation or involvement may develop in an individual. The approach is characterized as a motivational one. It uses the existing motivational language to explain work alienation and involvement for two basic reasons. First, theories of human motivation at work (Maslow, 1954; Lawler, 1973) are generally advanced to explain all work behavior, and alienation and involvement at work should not be considered exceptions. Second, the fact that the existing motivational constructs can adequately and parsimoniously explain work alienation phenomena lies hidden in many of the sociological and psychological formulations discussed earlier. Thus, a clearer motivational formulation of the phenomena is needed to bring this fact to the surface. In addition to the use of motivational language, the present approach is characterized by an emphasis on the following considerations.

1. In defining the state of work alienation, the approach limits itself to the analysis of

the behavioral phenomenon at the individual level. It identifies the state of alienation with a cognitive belief state of the individual. As a cognitive state, work alienation becomes conceptually distinct from many associated covert feelings (affective states of the individual expressed in terms of satisfaction or dissatisfaction experienced on the job) and overt behavior (job participation, assuming responsibility, etc.).

2. The approach emphasizes that the state of work alienation must be clearly distinguished from its causes (antecedent conditions) and its effects (consequent conditions). It considers the phenomenon to be caused by two sets of events, historical and contemporaneous. The approach also stresses that the cognitive state of work alienation has significant effects on subsequent job behavior and job attitudes.

3. The present approach can integrate and adequately explain the different types of alienation at work suggested by sociologists within its own framework.

It may be argued that the concepts of alienation and involvement should not be reduced to a single dimension, since they represent two distinct types of behavioral phenomena. The phenomenon of alienation has been described by sociologists at the collective level (alienation of labor, alienated society, etc.), whereas the phenomenon of involvement has been identified by psychologists at the individual level (involved worker). A closer scrutiny of the issue, however, reveals that even when sociologists describe the concept of alienation at the collective level, they try to explain the phenomenon in terms of psychological states of the individual. A number of empirical sociological studies on alienation (Blauner, 1964; Seeman, 1971; Shepard, 1971) attest to this fact. If one considers the phenomena of both alienation and 'involvement' to be states of the individual, it would be more parsimonious to consider both concepts as representing a single dimension than to consider them as independent dimensions.

In the following description of the present approach, the above-mentioned characteristics are highlighted.

Definitions of the Concepts

In the present approach, work involvement is viewed as a generalized cognitive (or belief) state of psychological identification with work insofar as work is perceived to have the potentiality to satisfy one's salient needs and expectations. Likewise, work alienation can be viewed as a generalized cognitive (or belief) state of psychological separation from work insofar as work is perceived to lack the potentiality for satisfying one's salient needs and expectations. Thus, the degree of involvement at work should be directly measured in terms of individual's cognition about his or her identification with work. The individual's identification with this work, however, depends on two things: the saliency of his or her needs (both extrinsic and intrinsic) and the perceptions he or she has about the need-satisfying potentialities of work.

Defining the concepts in this way has some implications for their measurement. If job involvement and alienation are viewed as cognitive states of an individual, they cannot be measured with the existing instruments (Blauner, 1964; Lodahl & Kejner, 1965; Saleh & Hosek, 1976; Shepard, 1971). Most of these instruments combine some measures of the cognitive state of alienation with some measures of its presumed causes and effects. For example, the most widely used instrument, developed by Lodahl and Kejner (1965), contains not only items that reflect the cognitive state of involvement ("I live, eat, and breathe my job") but also items that reflect both antecedent and consequent feeling states and behavioral tendencies ("I feel depressed when I fail at something connected with my job" or "I will stay overtime to finish a job, even if I am not paid for it"). Because of such built-in ambiguities in the existing instruments, the data these instruments yield are often hard to interpret. Future research efforts should attempt to develop more unambiguous measures of job involvement that reflect only the nature of the cognitive state of psychological identification with work. For instance, items such as "I live, eat, and breathe my job," "I am very much involved in my job," "The most im-

portant thing that happened to me involved my work," and so on tend to reflect the individual's awareness of work identification without measuring his or her need states (antecedent conditions) or overt behavioral tendencies (consequent conditions). These kinds of items have construct validity and therefore are more desirable measures of the cognitive state of job involvement. One can also use graphic techniques or the semantic differential format (Osgood, Suci, & Tannenbaum, 1957) to measure job involvement on dimensions such as involved–noninvolved, important–unimportant, identified–separated, central–peripheral, and so forth. Besides being less confusing with regard to assessing the cognitive states of involvement and alienation, such measures of job involvement that have construct validity seem to be better suited for cross-cultural and comparative research than are the existing measures, because the latter tend to mainly include and heavily emphasize items on intrinsic-need satisfaction. For groups of people who do not consider intrinsic needs (autonomy, control, etc.) to be the guiding forces in their lives, the existing measures with an emphasis on intrinsic needs cannot truly reflect their job involvement.

Conditions of Job Involvement

A schematic representation of the present motivational approach to job involvement, its causes, and its effects is presented in Figure 1. As can be seen in Figure 1, an individual's behavior and attitudes exhibited both on and off the job are a function of the saliency of need states within him or her. At any given moment, the need saliency within the individual depends on the prior socialization process (historical causation) and on the perceived potential of the environment (job, family, etc.) to satisfy the needs (contemporary causation). The cognitive state of involvement as a by-product of need saliency also depends on the nature of need saliency as historically determined through the socialization process and on the perceived potential of the environment to satisfy the needs. In the context of job involvement, an individual's belief that he or she is work involved or job alienated depends on whether the work is perceived to have the potential for satisfying his or her salient needs. The saliency or the importance of different needs for the individual is determined by the individual's past experiences with groups of which he or she was a member (socialization process) and with jobs that he or she has held. Different groups of people are influenced by different cultural, group, and organizational norms, and thus they tend to develop different need structures or to set different goals and objectives for their lives. For example, the work-motivation literature suggests that the sources of work involvement for managers within any organization may be very different from those for the unskilled laborers because of differences in the need saliencies of the two groups. Managers may value more autonomy and control in their jobs, whereas the unskilled laborers may attach greater importance to security and to sense of belonging in their jobs. Such value differences stem essentially from past socialization processes and from the influence of the norms of the groups to which the workers belong.

In a recent study, Kanungo, Gorn, and Dauderis (1976) demonstrated that because of differences in the socialization process, francophone and anglophone managers exhibit different patterns of need saliency at work. For instance, security and affiliation needs seem to have greater saliency for francophone as compared to anglophone managers, whereas autonomy and achievement needs tend to have greater saliency for anglophone as compared to francophone managers. The salient needs tend to determine the central life interests of the individuals. On the job, the saliency of a need in an individual may be reinforced when the person finds that through job behavior he or she is capable of meeting the needs. His or her perception that the job is capable of satisfying his or her important needs will make the individual devote most of his or her available energy to the job. The worker will immerse himself or herself in the job, and the feedback from his or her job behavior will lead the worker to believe that the job is an essential part of himself or herself or that he or she is job involved. If, how-

ever, the job is perceived by the individual as lacking in opportunities for the satisfaction of his or her salient needs, he or she will develop a tendency to withdraw effort from the job and thus become alienated from it. For the satisfaction of his or her salient needs, the person will redirect his or her energy elsewhere by engaging in various off-the-job activities or by engaging in various undesirable on-the-job activities.

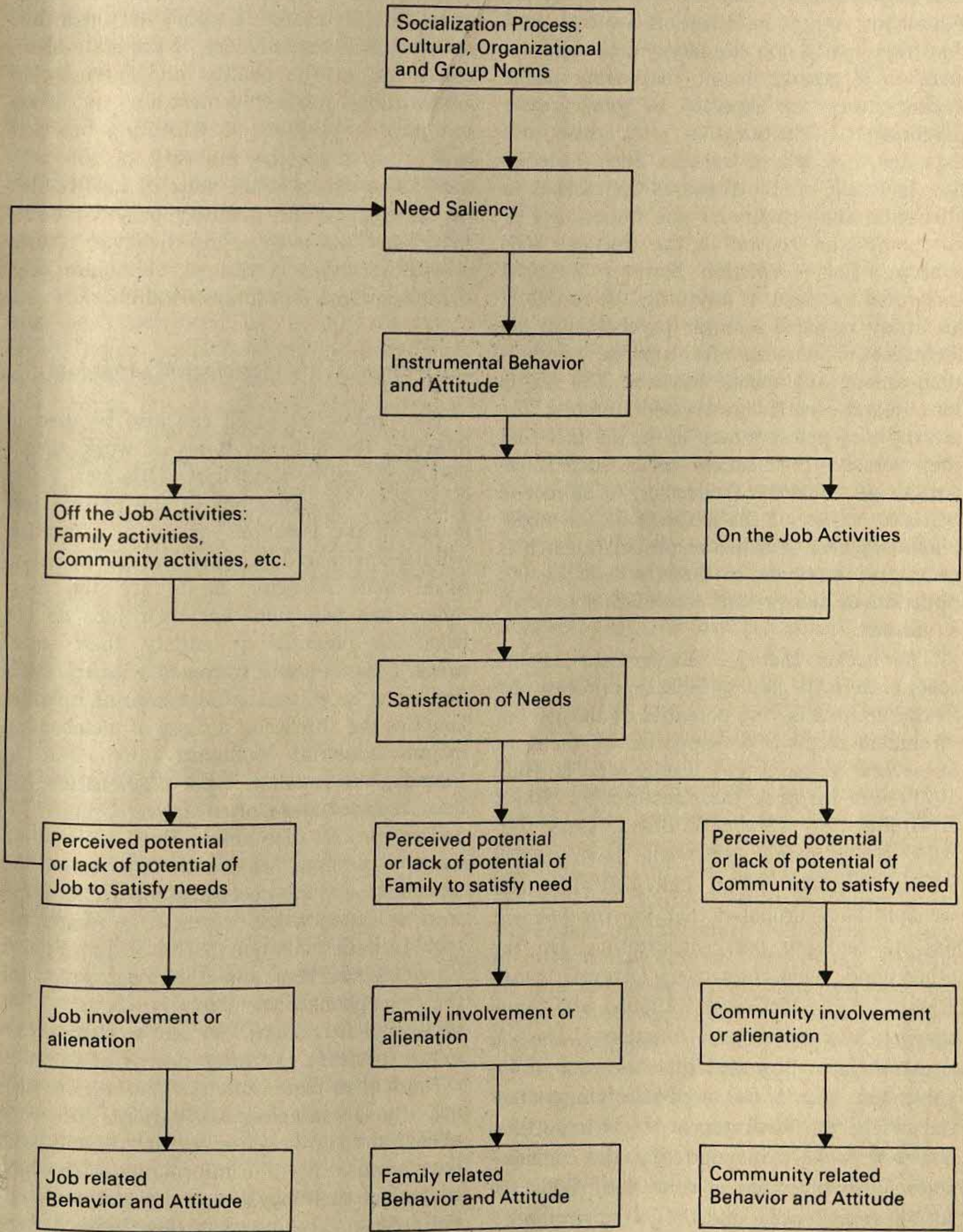

Figure 1. Schematic representation of the motivational approach to involvement and alienation.

A recent comparative study (Basu, 1976) of job involvement among francophone and anglophone workers has provided some indirect evidence in support of this motivational approach to work involvement. On the basis that anglophone workers are a product of the Protestant ethic socialization process and that they value job autonomy and achievement to a greater extent than francophone workers, they are expected to show greater psychological identification with their jobs than are francophone workers. Such a prediction is based on the previous approaches to alienation that emphasize the importance of autonomy and control in the worker's self-concept. This prediction however was not confirmed by Basu. If anything, the results of his study revealed stronger psychological job identification among francophone workers than among anglophone workers. The reason for greater work involvement among the francophone workers may lie in the fact that they perceive their salient needs, such as security and affiliation tendencies, to be met to a greater extent on the job than do the anglophone workers. Further empirical research is necessary, however, to directly test the implication of the present formulation in work situations.

The notion that job involvement has its roots both in the past socialization process and in the need-satisfying potential of the job environment seems to be supported by the work of several researchers (Rabinowitz & Hall, 1977). For instance, those researchers (Blood & Hulin, 1967; Hulin & Blood, 1968; Lodahl, 1964; Siegel, 1969) who have studied job involvement as an individual-difference variable have proposed that job involvement has its roots in past socialization. On the other hand, those researchers (Argyris, 1964; Bass, 1965; McGregor, 1960) who have studied involvement as a function of the job situation have proposed that the root of involvement lies in the need-satisfying potential of the job environment. It is important, however, to keep in mind that the cognitive state of involvement is caused by both the socialization process and the job environment. Future studies should be directed toward assessing the relative contributions of each of these causes of job involvement.

Figure 1 also suggests that a cognitive state of job involvement will have significant influence on job behavior and job attitudes. Several interesting possibilities present themselves in this regard. It would be worthwhile to investigate the influence of the state of involvement on the quality and intensity of job attitudes. Job involvement does not necessarily cause positive job attitudes, but perhaps does affect the intensity of job attitudes. The effects of the state of involvement on the quality and quantity of job productivity and on membership behavior (turnover, absenteeism, tardiness, etc.) also need to be investigated in future research.

Integration of the Sociological Approach

The present approach can also be used to interpret the different types of work alienation suggested by sociologists (Blauner, 1964; Seeman, 1959). In terms of the present formulation, the isolation variant of alienation will be experienced by those individuals whose social and belonging needs are the most salient and who find that their jobs do not have the potential to satisfy their social needs. Blauner seems to concur with this position when he states that the state of isolation "implies the absence of a sense of membership in an industrial community" (p. 24). In Canada, the isolation type of alienation has been reported more often among French Canadian workers than among English Canadian workers, perhaps because in the case of the former group, the necessary conditions for a state of isolation are present to a greater extent (salient affiliation needs of the French Canadian workers and their perception of the anglophone ownership of industry). For very similar reasons, female workers often may experience a greater degree of isolation at work than male workers. The normlessness and the meaninglessness variants of work alienation can be observed in persons who have a salient need for information with which to predict their physical and social work environments. Finding that their jobs do not provide the necessary information, they may

develop beliefs about the meaninglessness of their jobs. Educated and skilled workers may have a stronger need for information than lesss educated and unskilled workers. Hence the former group of workers may have a stronger tendency to develop beliefs about the meaninglessness of work than the latter group. Perhaps for similar reasons, the alienation of intellectuals tends to be of the meaninglessness variety (Seeman, 1959; Mills, 1951). The powerlessness type of alienated state may be experienced by individuals who have salient ego needs such as the need for autonomy, the need for control, and the need for self-esteem but who find the job environment incapable of satisfying them. Finally, the self-estrangement variety of alienation may be experienced by people who have highly salient self-actualization needs, such as the need for achievement, and find their jobs limiting the realization of their potentialities. Thus, from a motivational standpoint, the different types of alienation suggested by sociologists represent the same cognitive state of separation from an environmental entity and are different only in the sense that they are caused by the different saliency structure of needs in the individuals.

Some Major Differences Between the Present and the Earlier Approaches

At this point, it may be useful to compare and to highlight a few important differences between the present conceptualization and earlier ones.

Although the definitions of involvement and alienation as cognitive states of identification with work resemble the way the concepts were defined by Lawler and Hall (1970), the former are different from the latter in one important respect. As discussed earlier, Lawler and Hall (1970) put exclusive emphasis on the job opportunities that meet a worker's need for control and autonomy as necessary preconditions to the state of job involvement. In fact, all earlier formulations (both sociological and psychological) seem to have followed this line of thinking. The present approach, however, suggests that job involvement does not necessarily depend on job characteristics that allow for control- and autonomy-need satisfaction. It emphasizes that workers have a variety of needs, some more salient than others. The saliency of the needs in any given individual is determined by his or her past socialization in a given culture (historical causes) and is constantly modified by present job conditions (contemporary causes). Different groups of individuals, because of their different socialization training or different cultural background, may develop different need saliency patterns. They may value extrinsic and intrinsic job outcomes (Lawler, 1973) very differently. One set of needs (e.g., growth needs such as self-esteem and autonomy) may be salient in one group of workers, but the same needs may not be salient in another group. This may result in different self-images in the two groups and, consequently, in different job expectations in the two groups. One group of workers that considers control and autonomy to be the core of their self-image may get involved in jobs that are perceived as offering opportunity for exercising control and autonomy, and they may become alienated from jobs that are perceived as providing little freedom and control. Such job characteristics, however, may not be the crucial considerations for another group (who may view security and social needs to be the core of their self-image) in the determination of their job involvement or alienation. That people do differ with respect to what constitutes the core of their self-concepts should not be overlooked. The developed societies of the West may make their citizens believe that all that counts in one's life is to have individual liberty and freedom. Workers belonging to these societies may feel therefore that working life is of little worth without freedom and control. In contrast, however, in the developing societies of the East, economic and social security often are considered more important to life than are freedom and control. Thus, workers in eastern societies may find work very involving if it guarantees such security, but may not care very much for freedom and control in their jobs. In these societies, people may value equality more than liberty as the guiding principle of working life. Rabinowitz

and Hall (1977) alluded to this possibility, but found no available research that examined "this lower-need-based form of job involvement" (p. 280).

In their attempts to increase job involvement among workers, the sociological (Blauner, 1964) and the psychological (Lawler & Hall, 1970) approaches have analyzed the work situation from the standpoint of *job design*, or the nature of the job. They have emphasized job characteristics such as the lack of variety in a job, mechanized and routine operations, strict supervision, and so on and their effects on the involvement of workers without any attempt to understand the nature and the saliency of needs in the workers. In presenting such a position, these authors have argued in favor of a universal prescription for increasing job involvement by designing jobs to provide greater autonomy and control to the workers. The prescription is of course based on the assumption that the needs for control and autonomy are the most salient needs in workers. This position can be contrasted with the approach that Taylor (1911) advocated in his principles of scientific management. In his pig-iron-loading experiment, he selected as his subject a physically strong individual who had salient monetary need. In selecting the right man for the job, he looked into the past training and abilities, the need saliency, and the job perceptions of the worker. Obviously, Taylor must have thought that these characteristics have a significant influence on a worker's job involvement. The approach advocated in this article does not make the assumption that the needs for control and autonomy are the most salient needs in all workers. Unlike previous approaches, the present approach suggests that job involvement can be best understood if we find out the nature of and the saliency of needs in workers as they are determined by prior socialization and present job conditions. The design of jobs and the determination of their extrinsic and intrinsic outcomes for the sake of increasing job involvement should be based on an understanding of worker needs and perceptions. The findings of Lawler and Hackman (1971) seem to support this position. According to them, "There is no reason to expect job changes to affect the motivation and satisfaction of employees who do not value the rewards that their jobs have to offer" (p. 52).

Previous approaches emphasized the distinction between work as instrumental and work as consummatory activity (the means to an end vs. the end in itself). The present approach considers work to be a set of job-related behaviors and attitudes and, like all behaviors and attitudes, work is considered to be instrumental in satisfying a variety of needs that a worker may have. All human behaviors stem from need states, and all human behaviors tend to be purposive and instrumental in obtaining goals or outcomes for the satisfaction of needs. Work behaviors and job attitudes should not be an exception to this rule.

In summary, the motivational approach to the study of alienation and involvement advocated in this article provides an integrative framework for future psychological and sociological research. Future research in the area should not only attempt to measure job alienation and job involvement as cognitive states but also attempt to relate such cognitive states to the antecedent conditions of need saliency in the individual and his or her job perceptions. Attempts should also be made to relate the cognitive states of alienation and involvement to the various affective states that accompany them and to their behavioral consequences. Using the motivational approach, future studies should explore the phenomena of alienation and involvement in areas other than work, such as in the family, in the community, and in other forms of leisure-time pursuits (as suggested in Figure 1). It would be of considerable interest to find out the reasons for alienation and involvement in these areas for different groups of people with different socialization training. It would also be of interest to see how involvement and alienation in one area influence the nature of such states in other areas. For instance, how does job involvement affect family involvement and vice versa? The widely accepted Marxian dictum that work alienation is the cause of all social maladies is something that clearly

needs empirical verification. These are some of the general issues that need exploration in the future, and it is hoped that the framework proposed here will help in such exploration.

Reference Note

1. Faunce, W. *Occupational involvement and selective testing of self-esteem.* Paper presented at the meeting of the American Sociological Association, Chicago, 1959.

References

Alderfer, C. P. *Existence, relatedness, growth: Human needs in organizational settings.* New York: Free Press, 1972.

Allport, G. W. *Pattern and growth in personality.* New York: Holt, Rinehart & Winston, 1961.

Argyris, C. *Integrating the individual and the organization.* New York: Wiley, 1964.

Bass, B. M. *Organizational psychology.* Boston: Allyn & Bacon, 1965.

Basu, K. S. *Job involvement: An analysis in a bicultural context.* Unpublished master's thesis, McGill University, Montreal, Canada, 1976.

Blauner, R. *Alienation and freedom: The factory worker and his industry.* Chicago: University of Chicago Press, 1964.

Blood, M. R., & Hulin, C. L. Alienation, environmental characteristics, and worker responses. *Journal of Applied Psychology,* 1967, *51,* 284–290.

Clark, J. P. Measuring alienation within a social system. *American Sociological Review,* 1959, *24,* 849–852.

Denise, T. C. The concept of alienation: Some critical notices. In F. Johnson (Ed.), *Alienation: Concept, term, and meanings.* New York: Seminar Press, 1973.

Dubin, R. Industrial workers' worlds: A study of the central life interests of industrial workers. *Social Problems,* 1956, *3,* 131–142.

Dubin, R., Champoux, J. E., & Porter, L. W. Central life interests and organizational commitment of blue collar and clerical workers. *Administrative Science Quarterly,* 1975, *20,* 411–421.

Durkheim, E. *De la division du travail social.* Paris: F. Alcan, 1893.

Festinger, L. A theory of social comparison processes. *Human Relations,* 1954, *7,* 117–140.

Gerth, H. H., & Mills, C. W. *From Max Weber: Essays in sociology.* New York: Oxford University Press, 1946.

Hall, D. T., & Schneider, B. Correlates of organizational identification as a function of career pattern and organizational type. *Administrative Science Quarterly,* 1972, *17,* 340–350.

Hall, D. T., Schneider, B., & Nygren, H. T. Personal factors in organizational identification. *Administrative Science Quarterly,* 1970, *15,* 176–190.

Herzberg, F. *Work and the nature of man.* Cleveland, Ohio: World Publishing, 1966.

Hulin, C. L., & Blood, M. R. Job enlargement, individual differences, and worker responses. *Psychological Bulletin,* 1968, *69,* 41–65.

Johnson, F. (Ed.). *Alienation: Concept, term, and meanings.* New York: Seminar Press, 1973.

Jones, E. E., & Gerard, H. B. *Foundations of social psychology.* New York: Wiley, 1967.

Josephson, E., & Josephson, M. R. Alienation: Contemporary sociological approaches. In F. Johnson (Ed.), *Alienation: Concept, term, and meanings.* New York: Seminar Press, 1973.

Kanungo, R. N., Gorn, G. J., & Dauderis, H. J. Motivational orientation of Canadian anglophone and francophone managers. *Canadian Journal of Behavioral Science,* 1976, *8,* 107–121.

Lawler, E. E. *Motivation in work organizations.* Belmont, Calif.: Wadsworth, 1973.

Lawler, E. E., & Hackman, J. R. Corporate profits and employee satisfaction: Must they be in conflict? *California Management Review,* 1971, *14,* 46–55.

Lawler, E. E., & Hall, D. T. Relationship of job characteristics to job involvement, satisfaction, and intrinsic motivation. *Journal of Applied Psychology,* 1970, *54,* 305–312.

Lodahl, T. M. Patterns of job attitudes in two assembly technologies. *Administrative Science Quarterly,* 1964, *8,* 482–519.

Lodahl, T. M., & Kejner, M. The definition and measurement of job involvement. *Journal of Applied Psychology,* 1965, *49,* 24–33.

Macquarrie, J. A theology of alienation. In F. Johnson (Ed.), *Alienation: Concept, term, and meanings.* New York: Seminar Press, 1973.

Marx, K. [Economic and philosophical manuscripts.] In, *Marx-Engels Gesamtausgabe* (Vol. 3). Berlin, Germany: Marx-Engels Institute, 1932. (Originally published, 1844.)

Maslow, A. H. *Motivation and personality.* New York: Harper, 1954.

McGregor, D. *The human side of enterprise.* New York: McGraw-Hill, 1960.

Mills, C. W. *White collar.* New York: Oxford University Press, 1951.

Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. *The measurement of meaning.* Urbana: University of Illinois Press, 1957.

Patchen, M. *Participation, achievement, and involvement on the job.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Rabinowitz, S., & Hall, D. T. Organizational research on job involvement. *Psychological Bulletin,* 1977, *84,* 265–288.

Saleh, S. D., & Hosek, J. Job involvement: Concepts and measurements. *Academy of Management Journal,* 1976, *19,* 213–224.

Schacht, R. *Alienation*. Garden City, N.Y.: Doubleday, 1970.

Seeman, M. On the meaning of alienation. *American Sociological Review,* 1959, *24,* 783–791.

Seeman, M. The urban alienations: Some dubious theses from Marx to Marcuse. *Journal of Personality and Social Psychology,* 1971, *19,* 135–143.

Shepard, J. M. *Automation and alienation: A study of office and factory workers*. Cambridge, Mass.: MIT Press, 1971.

Siegel, L. *Industrial Psychology*. Homewood, Ill.: Irwin, 1969.

Taylor, F. W. *Principles of scientific management*. New York: Harper, 1911.

Vroom, V. Ego involvement, job satisfaction, and job performance. *Personnel Psychology,* 1962, *15,* 159–177.

Vroom, V. *Work and motivation*. New York: Wiley, 1964.

Weissenberg, P., & Gruenfeld, L. W. Relationship between job satisfaction and job involvement. *Journal of Applied Psychology,* 1968, *52,* 469–473.

Received October 5, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 139–147

Between-Subjects Expectancy Theory Research: A Statistical Review of Studies Predicting Effort and Performance

Donald P. Schwab, Judy D. Olian-Gottlieb, and Herbert G. Heneman III
Graduate School of Business and Industrial Relations Research Institute
University of Wisconsin—Madison

A large number of between-subjects expectancy theory studies have correlated measures of employee motivation (force) to perform (consisting of perceptions regarding linkages among effort and performance, performance and outcomes, and the attractiveness of the outcomes) with measures of effort and performance. The results of these studies (i.e., variance explained in effort and performance) vary considerably. A statistical review of these studies was conducted to determine the extent to which variance explained (the dependent variable) was a function of various characteristics of the effort and performance and of the force-to-perform measures (the independent variables). There were 160 observations, derived from 32 studies. Using multiple regression, we found that variance explained in these studies was greater when (a) self-report or quantitative measures of effort and performance were used rather than evaluations of these variables by someone other than the subject; (b) 10–15 outcomes were included in the force measure rather than a greater or smaller number of outcomes; (c) outcome valence was numerically scaled with positive numbers only, and the scale values were described in terms of desirability rather than importance; and (d) the force measure either contained no assessment of expectancy or an assessment that confounded expectancy and instrumentality. These variables accounted for 42% of the variance in the results obtained in the studies reviewed. Some theoretical and research implications of these findings are discussed.

Expectancy theory (Vroom, 1964) is a process theory of work motivation that has received considerable theoretical and empirical attention. Although numerous variations of the theory have been proposed (Campbell & Pritchard, 1976), the core of all formulations is the proposition that motivation (force) to perform is a function of (a) the expectancy that changes in effort will result in changes in performance, (b) the instrumentality of performance changes for the attainment of outcomes (such as pay increases), and (c) the valences of the outcomes. Also, in most formulations, these variables are combined in a multiplicative fashion as follows: Force to perform = f[expectancy $\times$ Σ(instrumentalities $\times$ valences)].

A substantial amount of research has been stimulated by expectancy theory. Almost all of it has used a between-subjects design in which a measure of force to perform is correlated with a measure of effort or performance for a sample of individuals. Some reviewers have objected to this orientation and have suggested that a within-subject design would be more appropriate for testing the theory (e.g., Mitchell, 1974). To date, how-

A preliminary version of this article was presented at the meeting of the National Academy of Management, Orlando, Florida, August 1977. The authors gratefully acknowledge the comments and criticisms of Chris Berger, Gil Churchill, and Larry Cummings in earlier stages of the project.

Requests for reprints or for a list of the studies reviewed in this article should be sent to Donald P. Schwab, Graduate School of Business, University of Wisconsin, 1155 Observatory Drive, Madison, Wisconsin 53706.

ever, there has been almost no such research on the prediction of effort or performance.

As the body of empirical literature on this theory has proliferated, so also have reviews summarizing and interpreting that literature (e.g., Campbell & Pritchard, 1976; Connolly, 1976; Heneman & Schwab, 1972; House, Shapiro, & Wahba, 1974; Mitchell, 1974; Mitchell & Biglan, 1971; Wahba & House, 1974). Although some of these reviews are essentially uncritical descriptions of the theory and research findings (e.g., House et al., 1974), all have made at least some suggestions aimed at improving the conduct of future research.

The suggestions that have been made can be viewed from two related perspectives. One is primarily theoretical in orientation. From this perspective, reviewers can be viewed as encouraging their empirically oriented colleagues to test expectancy theory qua theory. Thus, as examples, Heneman and Schwab's (1972) admonition to include assessments of expectancy in force measures and Mitchell's (1974) plea to perform within-subject studies can be viewed as attempts to make the measures and procedures of empirical investigations isomorphic with the theory.

An alternative perspective is to view these suggestions as aimed at increasing the theory's ability to predict effort or performance. The focal point of such an orientation is in increasing variance explained rather than in the elegance of the theoretical formulation per se. The reviews mentioned here have generally made suggestions regarding the measurement of the constructs specified in the theory. These recommendations, however, have in every case been based on casual assessments of previous research; that is, none of the reviews have empirically determined if variance explained differs as a function of the measurement procedures used.

The present review involves a statistical analysis of the results obtained in previously published between-subjects studies of expectancy theory. Studies included in our statistical analysis were those that examined the amount of variance explained in measures of effort or performance by measures of force to perform. These variance-explained values reported in the previous studies became our dependent variable. Independent variables chosen for the statistical analysis were various characteristics of (a) the measures of performance or effort and (b) the force measures that were used in the previously published research. Thus, the present study aims to account for variation in the strength of the relationship observed in other studies between force to perform and performance or effort as a function of characteristics of the effort, performance, and force-to-perform measures.

Three issues were examined regarding characteristics of the performance and effort measures. The first of these involved distinguishing between use of effort and use of performance as criterion measures. The conception of force is clearly aimed at predicting the effort expended toward some behavior. The actual behavior (job performance in the case at hand) is seen as a multiplicative function of force to perform and ability to perform (Vroom, 1964). Thus, from a theoretical perspective, one might argue, *ceteris paribus*, that more variance would be explained in studies using effort as the dependent measure than in studies using some measure of performance.

There is, however, reason to suspect that measures of effort might not be more predictable than measures of performance. Specifically, as Campbell and Pritchard (1976) have stated, "Organizational psychology is without any clear specification of the meaning of effort and consequently there is no operationalization of the variable that possesses even a modicum of construct validity" (p. 92). Thus, one might anticipate difficulty in predicting effort from force on pure measurement grounds.

A second dependent variable issue pertained to whether the measure was internal (self-reports) or external (others' ratings and rankings or productivity indexes) to the subject. Mitchell (1974) has argued that internal measures are more appropriate because effort is difficult to observe and hence measure externally. Moreover, it is generally agreed that when self-ratings are used simultaneously

for both the dependent and the independent variables, the resultant correlations are inflated. Thus, we hypothesized that more variance would be explained when self-ratings were used to measure effort or performance than when alternative measures were employed.

A third issue regarding the dependent variable was concerned with performance measures that were external to the subject. These measures have typically been ratings or rankings of performance provided most often by the subject's work supervisor. There have also been a number of instances in which investigators had access to quantitative productivity measures (units produced, sales volume, etc.). We hypothesized that more variance would be explained with the latter measures because of the reliability problems that frequently occur with performance ratings or rankings.

It was also possible to examine a number of issues dealing with characteristics of the force measure used. One important issue had to do with the number of outcomes (i.e., consequences of performance) assessed. Both Heneman and Schwab (1972) and Mitchell (1974) have suggested that a theoretical orientation would require that a large number of such outcomes be included in any test of the theory. This would be necessary to ensure that all outcomes of potential relevance to subjects are included in the measure of force to perform. No harm would come from this strategy because outcomes of no importance (with zero valence) would fall out of the force equation. For example, an outcome such as *improved recreational facilities* might not be important to most subjects. They would be expected to assign such an outcome zero valence, and hence it would not enter their force-to-perform equations. Nevertheless, including such an outcome would presumably enhance the force-to-perform estimates of those subjects for whom this outcome has some nonzero valence.

Again, however, psychometric issues potentially conflict with theoretical precision. Outcomes of little importance to subjects may contribute to unreliability of measurement and hence to reduced predictability. The potential conflict between theoretical and psychometric considerations suggests that a nonlinear relationship between number of outcomes and variance explained might be hypothesized. Up to a certain point increases in the number of second-level outcomes may increase variance explained because the marginal increment in total valence is large relative to the reliability decrement. We hypothesized, however, that beyond some point the marginal increment would be offset by the reliability decrement, resulting in a reduction in the variance explained.

The data permitted the examination of two issues involving the measurement of valence of second-level outcomes. One of these issues has to do with the verbal anchoring of the valence measures. Vroom (1964) defined valence as indicating anticipated satisfaction with, or desirability of, second-level outcomes. Both Connolly (1976) and Mitchell (1974) noted that many studies have anchored their valence scales with *importance,* which potentially represents an alternative construct. Connolly (1976) argued that unless the use of importance can be justified in variance-explained results, "There is a good argument for returning to the original conception of valence as anticipated satisfaction, or a close analog such as attractiveness, desirability, or anticipated utility" (p. 40). The second issue regarding valence has to do with the numerical anchors used. Again Mitchell (1974) has argued that theoretical purity requires that the anchors range from positive to negative values instead of using just the positive values that are frequently reported. Both of these valence issues (importance vs. desirability and positive to negative vs. positive only) were examined in the present study, although directional hypotheses were not specified a priori.

The final issues considered in the present study pertain to the measurement of expectancy. In their review Heneman and Schwab (1972) pointed out that most of the initial studies failed to measure expectancy at all or confounded expectancy with measures of instrumentality. They urged that future research include unconfounded mea-

sures of expectancy. This plea has apparently been heeded because a number of investigations have now been reported that include expectancy measures unconfounded with instrumentality. Moreover, Campbell and Pritchard (1976) concluded that measures of expectancy, considered singularly, tend to be positively correlated with measures of effort and performance. Thus, it was hypothesized that variance-explained values would be greater when the force measure included an expectancy term than when it did not. It was also hypothesized that more variance would be explained when this measure was not confounded with instrumentality. A related issue was the measurement of expectancy in those studies that included an unconfounded measure. Vroom (1964) defined expectancy as the subjective probability that an outcome (e.g., performance) will follow a specified level of effort. Thus, a theoretically correct measure of expectancy would assess it in likelihood terms, although in a number of studies it has been measured in alternative ways (e.g., having subjects compare the importance of personal effort relative to other potential determinants of performance; Schwab & Dyer, 1973). Following this theory we hypothesized that more variance would be explained when expectancy was measured in subjective-probability or likelihood terms.

Method

Dependent Variable and Population

The dependent variable was the amount of variance explained in a measure of effort or performance by a measure of force.[1] These values were easily obtained in studies that reported results in correlational terms by computing coefficients of determination (r^2 or R^2). The R^2 values were corrected for number of independent variables using Nunnally's (1967) correction formula. In two studies the results were not directly presented in correlational terms, but it was possible to derive the correlation from the information presented (Lawler, 1966; Turney, 1974).

A number of decisions about choice of the dependent variable were made. Many studies reported a variety of analyses; for example, performance might be correlated with a number of alternative force formulations. In these instances, data were used only from the model that most closely approximated the multiplicative force model specified earlier.[2] In addition, we included only the force measure that used the total number of outcomes assessed in the study. In the four studies that used cross-lagged correlation analysis (Kopelman & Thompson, 1976; Lawler, 1966; Lawler & Suttle, 1973; Sheridan, Slocum, & Richards, 1974), only the predictive results (force at time 1, effort or performance at time 2) were used. In all instances only one model of force was included from each study. However, an observation corresponding to each relationship between the force measure and alternative measures of effort or performance was included from those studies that had multiple measures of effort or performance.

A total of 32 published studies (using effort or performance measures as criteria) were found in which the results were reported in such a manner that a variance-explained value was presented or could be derived. Using the decision rules specified previously, a total of 160 observations were extracted from the 32 studies. For analysis purposes we viewed this $N = 160$ as the population of variance-explained values in between-subjects studies.

Independent Variables

The studies were reviewed to obtain the necessary independent variable information. Since the information was relatively straightforward, little ambiguity occurred in reviewing the studies and coding the data. The following independent variables were used in the present study: (a) whether effort or performance was measured, (b) whether the effort/performance measure was self-reported or externally assessed, (c) whether the externally assessed measure was based on objective data (e.g., sales volume and productivity) or subjective appraisal, (d) number of second-level outcomes (trichotomized into categories of approximately equal numbers of observations consisting of 1–9 outcomes, 10–15 outcomes and 16 or more outcomes), (e) whether valence was scaled positive to negative or only positive, (f) whether the verbal anchor for valence was importance or desirability,[3] (g) whether an expectancy

---

[1] The dependent variable used in the present study resulted in the exclusion of certain well-known expectancy studies (e.g., Georgopolous, Mahoney, & Jones, 1957; Porter & Lawler, 1968) because their results could not be cast into a variance-explained format. Also, negative force–performance (or force–effort) relationships were coded zero in the analysis.

[2] One exception to the decision rule was made for the Oliver (1974) study. He did not report results for the multiplicative model alone. Thus, the variance-explained estimate recorded for his study consisted of the multiplicative and additive models combined.

[3] In coding the verbal valence anchors, Graen's (1969) essential–unnecessary categories were coded

Table 1
Frequency Distribution of Independent Variables

Variable	f
Dependent variable	
Productivity	28
Self-report effort	29
Self-report performance	15
Other-report effort	13
Other-report performance	75
Number of outcomes	
≤ 9	46
10–15	61
≥ 16	53
Valence characteristic	
Importance	
Negative–positive	1
Positive only	58
Desirability	
Negative–positive	72
Positive only	29
Expectancy characteristic	
Unconfounded	
Likelihood estimate	27
Other	69
Confounded	51
None	13

Note. N = 160.

measure was included in the force measure, (h) whether the expectancy measure was unconfounded with instrumentality, and (i) whether expectancy was measured as a likelihood estimate.

Analysis

All independent variables were dummy coded (Cohen, 1968). The frequency with which each category appeared in the studies reviewed is shown in Table 1. Table 1 shows, for example, that 28 of the observations had an externally assessed productivity measure, that 29 had a self-report measure of effort, and so forth. Frequencies within each general category (dependent variable, number of outcomes, valence characteristics, and expectancy characteristics) sum to the total ($N = 160$) because they are made up of mutually exclusive and exhaustive subcategories. The mean variance explained was calculated for each of the categories shown in Table 1. In addition, the variance-explained values were regressed on various combinations of the dummy categories. This analysis was performed to assess how much of the variability in the results of previous research could be accounted for by the procedural characteristics of the studies. All multiple coefficients of determination reported in subsequent tables are significant ($p < .05$). However, significance levels are not reported in the tables because we view this analysis as describing the population of between-subjects, variance-explained estimates of effort and performance. Some inferential implications of this study are considered in the discussion.

Results

The first analysis involved an examination of the variance-explained values from the 160 observations in terms of the characteristics of the variables used to measure effort and performance. Table 2 reports the average variance explained between measures of force and the five classifications of the dependent variables employed in the studies reviewed. As hypothesized, force and self-report measures of effort and performance are more highly related than are force and others' assessment of effort or performance. For example, measures of force account for 10% of the variance on the average in self-report measures of performance and for 7% in performance measures assessed by others. Table 2 also shows, as hypothesized, that quantitative measures of productivity are more predictable than others' ratings or rankings of performance or effort. On the other hand, effort was less predictable than performance only for others' ratings and rankings. Overall, method of categorizing the measures of effort and performance accounted for 8% of the variability in the variance-explained values.

Table 2
Variance Explained as a Function of Type of Effort or Performance Measure

Dependent variable	M variance explained
Productivity	.13
Self-report effort	.13
Self-report performance	.10
Other-report effort	.03
Other-report performance	.07
R^2	.08

as important–unimportant. Coded as desirable–undesirable were good–bad (Hackman & Porter, 1968; Lied & Pritchard, 1976; Matsui & Terai, 1975), attractive–unattractive (Pritchard & Sanders, 1973), and preferences among pairs of outcomes (Sheridan et al., 1974).

Table 3
Variance Explained as a Function of Number of Outcomes

Number of outcomes	M variance explained
≤ 9	.08
10–15	.14
≥16	.05
R^2	.12

Table 3 shows the variance that force measures account for in measures of performance and effort as a function of the number of outcomes assessed. As hypothesized, highest average variance is explained in studies with an intermediate number of outcomes (14%), somewhat less variance is explained in studies with 9 or less outcome (8%), and the least variance is explained in studies with 16 or more outcomes (5%). Twelve percent of the variability in the results of the studies reviewed is accounted for by this categorization of the number of outcomes.

Preliminary analysis found that the numerical (negative to positive vs. positive only) and verbal (desirability vs. importance) anchoring procedures for valence measures were not independent (see Table 1). As a consequence, four categories were established for valence, as shown in Table 4. It can be seen that desirability scalings on the average have yielded higher variance-explained estimates than have importance scalings. Thus, Vroom's (1964) original definition receives some empirical support in terms of the verbal anchors. On the other hand, Table 4 shows that positive-only anchors, especially when combined with desirability anchors, have resulted in the highest average variance explained. All told, the scaling of valence accounts for 10% of the variability in the results obtained in the expectancy research reviewed here.

Table 4
Variance Explained as a Function of Valence Characteristics

Characteristic	M variance explained
Importance	
Negative–positive	.05
Positive only	.08
Desirability	
Negative–positive	.07
Positive only	.16
R^2	.10

Table 5 shows average variance explained as a function of the measurement of expectancy. Contrary to the hypothesis, greatest average variance is explained in studies that did not include a measure of expectancy (12%) or that confounded this measure with instrumentality (14%). Moreover, slightly less variance is explained in studies that used a likelihod estimate (5%) than in those that used alternative procedures (6%) among studies that did include an unconfounded measure. Twelve percent of the variability in the results of the studies reviewed is accounted for by the categorization of the procedures used to assess expectancy.

A final equation was generated by regressing the variance explained in previous expectancy studies on all of the independent variables simultaneously. The signs of the regression coefficients in this equation were the same as in the equations generated to obtain R^2 values in Tables 2–5. The R^2 for this last equation was .42. Thus, 42% of the variability in the results of the studies reviewed is accounted for by the categorizations of the dependent variables and force measures.

Discussion

At the outset it is important to recognize that our analysis was necessarily constrained by the measures and procedures used in the studies reported in the literature. Thus, as an example, Heneman and Schwab (1972) called for comparisons of results obtained using additive versus multiplicative combinations of force measures. Schmidt's (1973) criticism of multiplicative analyses aside, the present study was forced to consider multiplicative models because of those studies included in this review, only the Dyer and Weyrauch (1975), Oliver (1974), Pritchard and Sanders (1973), and Schwab and Dyer (1973) studies reported additive (or additive plus

Table 5
Variance Explained as a Function of Expectancy Characteristics

Characteristic	*M* variance explained
Unconfounded measure	
Likelihood estimate	.05
Other	.06
Confounded measure	.14
None	.12
R^2	.12

interactive) combinations of all force components in addition to the multiplicative mode of combination.

An additional issue that could have been investigated, but was not, is the distinction between models that contrast results obtained using so-called intrinsic versus extrinsic outcomes. House et al. (1974), Wahba and House (1974), and Mitchell (1974) have all urged that such distinctions be made, and, indeed, a number of studies have purportedly done so (e.g., Mitchell & Albright, 1972; Oliver, 1974). However, such distinctions seem arbitrary in view of Dyer and Parker's (1975) demonstration that one social scientist's extrinsic outcome is another's intrinsic outcome and vice versa. As a consequence, we considered only the models in each study that included all outcomes.

Nevertheless, the issues that were investigated in the present review accounted for a substantial portion of the variance in the results of between-subjects expectancy theory studies designed to predict effort or performance. This was accomplished by categorizing several characteristics regarding the operationalization of the dependent variable and of force to perform. We found that self-report measures were more highly related to measures of force than were measures provided by other evaluators. This finding has been observed by other reviewers and probably reflects spurious method covariation. Objective measures of performance were also associated with greater variance explained than were measures obtained from other evaluators. There are at least two possible explanations for the higher predictability of objective measures. One possibility is that quantitative measures are more reliable than measures provided through an appraisal process. An alternative explanation has to do with the possible boundary conditions of the theory. Dachler and Mobley (1973) have suggested that the theory may only be predictive in situations in which outcomes are objectively linked to behaviors. It may be that in situations in which performance is objectively measured, the organization is more likely to use contingent reward systems (particularly those involving monetary rewards).

The composition of the force measure used was also related to the results obtained by previous researchers. Studies that used 10–15 second-level outcomes obtained stronger relationships between force and performance or effort than did studies that used either fewer or more outcomes. Additionally, studies that scaled valence only positively and that used desirability–undesirability verbal anchors resulted in more variance explained than alternative formulations of valence. Studies that did not measure expectancy at all or that confounded expectancy with instrumentality measures obtained stronger results than those that measured expectancy in a more theoretically correct fashion.

It is obvious from these results that maximum variance explained in between-subjects predictions of performance or effort has not been obtained by making force measures adhere to the theory. Indeed, every finding regarding the measurement of force could be interpreted as contrary to the theory except for the verbal anchoring of valence. Models have yielded the strongest results without a theoretically appropriate expectancy measure, with a moderate number of second-level outcomes, and with valence scaled only in a positive direction.

It is tempting to infer from these findings the likely results that would be obtained if one were to conduct a between-subjects study of performance or effort based on expectancy theory. The major probable constraint on inference, however, stems from the fact that multiple observations were taken from many of the studies reviewed. This clustering of observations within studies results in the probable underestimation of the standard

error of estimates and hence in the overestimation of the corresponding *F* values (Kish, 1957). Unfortunately, since the nonindependence of dependent values within clusters, if any, is confounded with the impact of the independent variables investigated, we know of no way to identify the magnitude of the problem and still retain all 160 observations.

We attempted to obtain some information regarding the appropriateness of generalization by replicating the analyses performed on the population in two independent subsamples drawn from the population. These samples were obtained by randomly choosing one observation per study, subject to the constraint that an observation could appear only once in the two samples. This procedure resulted in subsamples of $n = 31$ and $n = 27$ (six studies had only one observation). Variance explained was used as a dependent variable, as were two transformations aimed at generating a dependent distribution approximating normality. The first was Fisher's r to z transformation, and the second was a transformation derived according to procedures suggested by Hinkley (1977).

Generally speaking, the direction of results from these analyses was similar to the direction of results obtained on the population of observations. Self-report dependent variables were more predictable than others' reports in both samples, as in the population. Productivity was more predictable than others' reports in one sample, but the two were about equally predictable in the other. The intermediate number of outcomes was associated with highest average variance explained in one sample. In both samples, as in the population, lowest variance explained on the average occurred in the category with 16 or more outcomes. Also as in the population, valence scaled only positively resulted in greater variance explained in both samples. Moreover, in both samples the average variance explained was greater in studies that did not measure expectancy or that confounded this measure with instrumentality.

However, none of the coefficients of determination generated on the variance-explained values or on the transformed values were statistically significant ($p < .05$) in either sample. This lack of significance is probably due to the small sample sizes that were necessary to achieve independence of observations, as well as due to the large numbers of independent variables (relative to sample size). Thus, inferences drawn about the probable outcomes of future research findings from the results obtained here must be made cautiously.

Despite these qualifications, there is a nagging suspicion that expectancy theory overintellectualizes the cognitive processes people go through when choosing alternative actions (at least insofar as choosing a level of performance or effort is concerned). The results of the present review are consistent with this suspicion. At the very least, whether for theoretical or measurement reasons, our results indicate that complicated measures of force have not aided prediction in between-subjects investigations.

References

Campbell, J. P., & Pritchard, R. D. Motivation theory in industrial and organizational psychology. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 1976.

Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, *70*, 426–443.

Connolly, T. Some conceptual and methodological issues in expectancy models of work performance motivation. *Academy of Management Review*, 1976, *1*(4), 37–47.

Dachler, H. P., & Mobley, W. H. Construct validation of an instrumentality–expectancy–task-goal model of work motivation: Some theoretical boundary conditions. *Journal of Applied Psychology*, 1973, *58*, 397–418. (Monograph)

Dyer, L. D., & Parker, D. F. Classifying outcomes in work motivation research: An examination of the intrinsic–extrinsic dichotomy. *Journal of Applied Psychology*, 1975, *60*, 455–458.

Dyer, L. D., & Weyrauch, W. MBO and motivation: An empirical study. *Academy of Management Proceedings*, 1975, pp. 134–136.

Georgopoulos, B. S., Mahoney, G. M., & Jones, N. W. A path-goal approach to productivity. *Journal of Applied Psychology*, 1957, *41*, 345–353.

Graen, G. Instrumentality theory of work motivation: Some experimental results and suggested modifications. *Journal of Applied Psychology Monograph*, 1969, *53*(1, Pt. 2).

Hackman, J. R., & Porter, L. W. Expectancy theory predictions of work effectiveness. *Organizational Behavior and Human Performance*, 1968, *3*, 417–426.

Heneman, H. G., III, & Schwab, D. P. Evaluation of research on expectancy theory predictions of employee performance. *Psychological Bulletin,* 1972, *78,* 1–9.

Hinkley, D. On quick choice of power transformation. *Applied Statistician,* 1977, *26,* 67–69.

House, R. J., Shapiro, H. J., & Wahba, M. A. Expectancy theory as a predictor of work behavior and attitude: A re-evaluation of empirical evidence. *Decision Sciences,* 1974, *5,* 481–506.

Kish, L. Confidence intervals for clustered samples. *American Sociological Review,* 1957, *22,* 154–165.

Kopelman, R. E., & Thompson, P. H. Boundary conditions for expectancy theory predictions of work motivation and job performance. *Academy of Management Journal,* 1976, *19,* 237–258.

Lawler, E. E., III. Ability as a moderator of the relationship between job attitudes and job performance. *Personnel Psychology,* 1966, *19,* 153–164.

Lawler, E. E., III, & Suttle, J. L. Expectancy theory and job behavior. *Organizational Behavior and Human Performance,* 1973, *9,* 482–503.

Lied, T. R., & Pritchard, R. D. Relationships between personality variables and components of the expectancy–valence model. *Journal of Applied Psychology,* 1976, *61,* 463–467.

Matsui, T., & Terai, T. A cross-cultural study of the validity of the expectancy theory of work motivation. *Journal of Applied Psychology,* 1975, *50,* 263–265.

Mitchell, T. R. Expectancy models of job satisfaction, occupational preference and effort: A theoretical, methodological, and empirical appraisal. *Psychological Bulletin,* 1974, *81,* 1053–1077.

Mitchell, T. R., & Albright, D. W. Expectancy theory predictions of the satisfaction, effort, performance, and retention of naval aviation officers. *Organizational Behavior and Human Performance,* 1972, *8,* 1–20.

Mitchell, T. R., & Biglan, A. Instrumentality theories: Current uses in psychology. *Psychological Bulletin,* 1971, *76,* 432–454.

Nunnally, J. C. *Psychometric theory.* New York: McGraw-Hill, 1967.

Oliver, R. L. Expectancy theory predictions of salesmen's performance. *Journal of Marketing Research,* 1974, *11,* 243–252.

Porter, L. W., & Lawler, E. E., III. *Managerial attitudes and performance.* Homewood, Ill.: Irwin, 1968.

Pritchard, R. D., & Sanders, M. S. The influence of valence, instrumentality, and expectancy on effort and performance. *Journal of Applied Psychology,* 1973, *57,* 55–60.

Schmidt, F. L. Implications of a measurement problem for expectancy theory research. *Organizational Behavior and Human Performance,* 1973, *10,* 243–251.

Schwab, D. P., & Dyer, L. D. The motivational impact of a compensation system on employee performance. *Organizational Behavior and Human Performance,* 1973, *9,* 215–225.

Sheridan, J. E., Slocum, J. W., Jr., & Richards, M. D. Expectancy theory as a lead indicator of job behavior. *Decision Sciences,* 1974, *5,* 507–522.

Turney, J. R. Activity outcome expectancies and intrinsic activity values as predictors of several motivation indexes for technical-professionals. *Organizational Behavior and Human Performance,* 1974, *11,* 65–82.

Vroom, V. H. *Work and motivation.* New York: Wiley, 1964.

Wahba, M. A., & House, R. J. Expectancy theory in work and motivation: Some logical and methodological issues. *Human Relations,* 1974, *27,* 121–147.

Received October 7, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 148–150

The Solzhenitsyn Finger Test: A Significance Test for Spontaneous Recovery

Edmund B. Coleman
University of Texas at El Paso

Spontaneous recovery is generally considered to be an unreliable phenomenon, but this is because of the fact that an inappropriate significance test has been used in past research. The research hypothesis states that spontaneous recovery is a monotonic function of time. Time is an ordered continuous variable; it flows steadily in one direction. The interaction test that has been traditional, however, is only appropriate for an unordered discrete variable. To illustrate this point, the study most frequently cited as evidence against spontaneous recovery was retested with significance tests appropriate for an ordered variable (specifically, the correlation coefficients tau and rho), and the findings were shown to be evidence for, not against, a recovery effect.

Spontaneous recovery in verbal learning is not widely considered to be a reliable phenomenon. In his general text on verbal learning, Jung's (1968) first sentence about the topic is, "Evidence regarding spontaneous recovery has been weak" (p. 118). Saltz (1971) wrote, "None of the studies . . . has reported evidence supporting the spontaneous recovery theory of proactive inhibition" (p. 218). Much the same evaluation is found in other texts and in reviews (Hall, 1971, p. 492; Hulse, Deese, & Egeth, 1975, p. 353; Keppel, 1968, p. 185).

But after an overall review of the research, Brown (1976) reached a different conclusion and suggested that the problem may be "that a more sensitive statistical or analytical measure may be required" (p. 336). The purpose of this article is to state Brown's suggestion in stronger language. The spontaneous recovery design needs a *correct* analysis; the interaction test that has been used for over 20 years does not fit the research hypothesis.

The most frequently cited study of spontaneous recovery (Koppenaal, 1963) can serve to illustrate the design and traditional analysis. Koppenaal used an A-B, A-C paradigm; that is, his subjects first learned an A-B list (*shiny–bitter*) and then an A-C list (*shiny–pretty*). There was also a control group who learned a single list. Figure 1 presents his findings, but the plot has been simplified by omitting data unrelated to the present argument. The traditional test for spontaneous recovery has been to test the mean square of the interaction between lists and time. Koppenaal's nonsignificant F of 1.30 and similar findings by Slamecka (1966) have been widely cited as evidence against spontaneous recovery. But given Brown's (1976) analysis of the overall evidence, the negative evaluation of spontaneous recovery findings by those in the field becomes unconvincing. A reexamination of the statistical logic appears in order.

The Solzhenitsyn Finger Test

In *August 1914* (1971/1972) Solzhenitsyn has his alter ego demonstrate to General Samsonov that the Russian front line is dangerously overextended with what could be called the Solzhenitsyn Finger Test, an important advance in instrumentation over the Interocular Traumatic Test (Edwards, Lindman, & Savage, 1963). Colonel Voro-

Requests for reprints should be sent to Edmund B. Coleman, Department of Psychology, University of Texas, El Paso, Texas 79968.

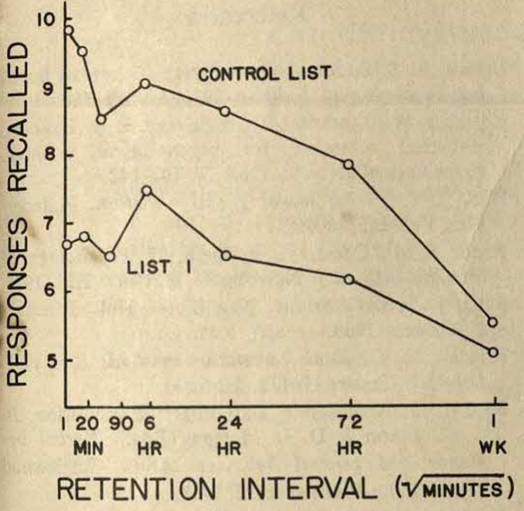

Figure 1. Number of correct responses for List 1 and the control list. The data are from Koppenaal (1963).

tyntsev, using his fingers as dividers, swings off six spans on the war map, showing Samsonov that his front-line troops are 6 full marching days ahead of staff headquarters.

Let us apply Colonel Vorotyntsev's finger test to the curves of Figure 1. The research hypothesis is that as time passes, the suppressed associations in List 1 spontaneously recover, thus causing its curve to converge toward the curve of the control list. Using your fingers as dividers, measure the difference between List 1 and the control list for the 1-minute interval. It will be about 3.2 responses. Now do the same for the 20-minute interval. It will be smaller—about 2.8 responses. If you continue for the longer and longer intervals, you will get smaller and smaller differences.

The charade has been pushed dangerously close to the point of insult. The point is that time is an ordered continuous variable; it only flows in one direction. Thus, the research hypothesis predicts an *ordered* decrease in differences; the significance test that has been in use for over 20 years, however, is *unordered*. Researchers have been using the interaction between lists and time, and that interaction is based on the unordered variance between the differences. The interaction mean square for Koppenaal's curves would have been exactly the same if the largest difference (3.2) had occurred at the second interval, the third interval, or at any other interval including the one for 1 week.

There are a number of powerful tests that fit an ordered hypothesis like the one for spontaneous recovery—the difference between correlation coefficients, the difference between slope coefficients, the linear component of the differences between the curves, and the like. Applying the more powerful tests to Koppenaal's data, however, would require estimation of certain terms and extensive justification of these estimations. A more economically presented test could be based on Kendall's (1962) tau. Since the test would use the degrees of freedom for the seven intervals instead of those for Koppenaal's 168 subjects, it would be somewhat overconservative; but perhaps that is a virtue in the present context.

For the seven intervals, the differences between the two curves are 3.2, 2.7, 2.15, 1.55, 2.2, 1.6, and .5. The research hypothesis (that as List 1 spontaneously recovers over time, the differences will become progressively smaller) can be expressed as 18 predictions: The difference between the curves at the 1-minute interval will be larger than the following six differences (all 6 predictions are correct). The difference at the 20-minute interval will be larger than the following five differences (all 5 predictions are correct). The difference at the 90-minute interval will be larger than the following four (3 predictions are correct and 1 is incorrect), and so on, as per Kendall (1962). Out of the 18 predictions, 15 conform to the research hypothesis. The ratio of 15 to 3 can be evaluated by Kendall's table of S values ($S = 18 - 3 = 15$) and is significant ($p < .015$, one-tailed).

An approach with the same underlying logic is to correlate the seven differences with their intervals ($\rho = .82$, $p < .025$). If one uses a more powerful test that replaces the degrees of freedom for the seven intervals with the degrees of freedom for Koppenaal's (1963) 168 subjects, the significance level will become more extreme. In short, the

finding most frequently cited as evidence against spontaneous recovery is, on the contrary, strong evidence in its favor.

In his Table 1, Brown (1976) listed 21 studies of spontaneous recovery. All 21 showed a trend toward recovery, but 8 of them (including Koppenaal's) reported a nonsignificant interaction. If the studies were retested with an ordered test, it is likely that several more of the negative reports would turn out to be evidence in favor of spontaneous recovery.

It is important to note that Type I errors are possible when the interaction is used to test an ordered function. In the case of spontaneous recovery, it is true, the opposite error is more likely to have been made. Because of the excessive weakness of the inappropriate test, which in turn led to a few incorrect negative evaluations, several characteristics of spontaneous recovery that are cornerstones of its theoretical relevance have not been vigorously pursued as research topics, for example, recovery over long intervals, recovery of specific suppressions, recovery of bonds not taught in laboratory lists, and others. Given the warranted power of an ordered test, perhaps such characteristics will appear in future research.

References

Brown, A. S. Spontaneous recovery in human learning. *Psychological Bulletin,* 1976, *83,* 321–338.

Edwards, W., Lindman, H., & Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review,* 1963, *70,* 193–242.

Hall, J. F. *Verbal learning and retention.* Philadelphia, Pa.: Lippincott, 1971.

Hulse, S. H., Deese, J., & Egeth, H. *Psychology of learning* (4th ed.). New York: McGraw-Hill, 1975.

Jung, J. *Verbal learning.* New York: Holt, Rinehart & Winston, 1968.

Kendall, M. G. *Rank correlation methods* (3rd ed.). London: Charles Griffin, 1962.

Keppel, G. Retroactive and proactive inhibition. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory.* Englewood Cliffs, N. J.: Prentice-Hall, 1968.

Koppenaal, R. J. Spontaneous recovery? *Journal of Verbal Learning and Verbal Behavior,* 1963, *2,* 310–319.

Saltz, E. *The cognitive bases of human learning.* Homewood, Ill.: Dorsey Press, 1971.

Slamecka, N. J. A search for spontaneous recovery of verbal associations. *Journal of Verbal Learning and Verbal Behavior,* 1966, *5,* 205–207.

Solzhenitsyn, A. I. [*August 1914*] (M. Glenny, trans.). New York: Farrar, Straus & Giroux, 1972. (Originally published, 1971.)

Received October 11, 1977

Psychological Bulletin
1979, Vol. 86, No. 1, 151–167

On the Nature of Taste Qualities

Donald H. McBurney and Janneane F. Gent
University of Pittsburgh

The concept of four basic tastes developed historically on the basis of a number of criteria. Modern evidence, largely electrophysiological, has led some investigators to reject the concept of basic tastes in favor of a taste continuum or multidimensional space. The present article reviews data that support the validity of the basic tastes. It is concluded that this concept, as well as the separate question of taste as an analytic or synthetic sense, is compatible with either of the two major positions on sensory coding in taste, labelled line theory and pattern theory. Taste may profitably be considered to comprise four distinct but interacting sensory modalities or submodalities analogous to the (other) skin senses.

Every introductory psychology text informs its readers that there are four taste qualities: sweet, salty, sour, and bitter. Such a statement may reflect the views of a majority of workers in taste at the present time, but it glosses over a long historical development and considerable current ferment (McBurney, 1974). On the one hand, those in the psychophysical tradition tend to accept the consensus of four taste qualities because of its obvious convenience and predictive usefulness. Those who are more physiologically oriented, however, tend to reject the notion of four taste qualities as naive. For example, Uttal (1973) wrote, "The psychophysical evidence continually seems to evoke the use of the four basic taste words. Yet . . . it is moot whether this must be interpreted as a reflection of the underlying biological mechanisms or of the evolved language of gustatory experience" (p. 603). Similarly, Schiffman and Erickson (1971) commented with respect to their psychophysical model of taste, "We would like to emphasize that the present model does not require, or support, the idea of taste primaries. That is, although the ordering of stimuli may, with some imagination, be seen as resulting in four rather indistinct stimulus groups . . . we do not suggest here that . . . any of these stimuli may be usefully described by primary tastes" (p. 631).

It is our purpose in this article to review the history of the development of the concept of taste qualities and to discuss the theoretical issues involved in relating the psychophysical and physiological approaches to understanding sensory coding in taste. We argue that the four taste qualities together exhaust the qualities of taste experience. We propose further that the four taste qualities, far from being arbitrary, may profitably be thought of as representing separate but interacting sensory systems in the same way that we think of the skin senses. Although the focus of the article is on the status of the qualities of taste, it is necessary to discuss the related issues of the nature of the stimulus dimension(s) for taste and the physiological mechanisms that code the dimension(s). The two major positions of sensory coding in taste are known as the labelled line theory and the across-fiber pattern theory. The labelled line theory holds that the quality information is carried by the fiber that is most responsive to a particular stimulus, whereas the pattern theory holds that the information consists in the relative firing rate of two or more fibers.

This research was supported by National Institutes of Health Grant 2 RO1 NS 07873. We thank L. M. Bartoshuk and L. E. Marks for helpful discussions.

Requests for reprints should be sent to D. H. McBurney, Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

Furthermore, the question of whether taste is analytic or synthetic in the way it responds to mixtures has become an issue in the debate over gustatory sensory coding. We argue that the question of the status of the four taste qualities is independent of the resolution of both the analytic–synthetic debate and the labelled line versus pattern theory debate.

It is necessary at this point to define taste, for the purpose of this article, as those sensations mediated by the taste buds. As such, the sense of taste is part of a perceptual system that involves all of the chemically sensitive nerves and end organs of the oral and nasal cavities that aid in the investigation of the chemical environment (cf. Gibson, 1966). Thus, taste is distinguished from flavor, which comprises other sensations mediated by olfaction, touch, temperature, and even vision. The distinction between taste and flavor has formed the basis of the vast bulk of the work in taste and is crucial to the understanding of the logical status of the taste qualities.

Relation of Psychophysical Data to Physiological Data

A fundamental problem in the analysis of a sensory system is to relate physiological and psychophysical data, that is, to develop what Brindley (1960) called *linking hypotheses:* "If a physiological hypothesis, i.e. a hypothesis about function that is stated in physical, chemical and anatomical terms, is to imply a given result for a sensory experiment, the background of theory assumed in conjunction with it must be enlarged to include hypotheses containing psychological terms as well as physiochemical and anatomical" (p. 145) ones.

The problem in taste research is that there is, at present, little agreement on the status of the psychological terms that are presumably to be explained by the physiological data by the way of linking hypotheses. Although some may wish not to be constrained by the categories of human taste experience, it is certainly true that terms like sweet and salty, which are unquestionably derived from human experience, are used in taste physiology in ways that appear definitive. Are the taste qualities primaries and, if so, in what sense of the word? Are they arbitrary terms borrowed in a naive way from our language? Or, are they simply arbitrary locations in a multidimensional perceptual space?

Uses of the Term Primary

The psychological distinctiveness of sweet, salty, sour, and bitter has tempted many to label these qualities *primary*. The term *primary* appears to have been borrowed from vision, and its application to any other sensory system unavoidably invites comparison with color perception. The term can be a source of confusion because it is used in a number of different ways. It is used variously to refer to (a) sensations that are psychologically (phenomenologically) distinct and that cannot be analyzed introspectively into two or more categories, such as the color yellow; (b) stimuli that may be mixed to produce sensations that are normally associated with a third stimulus, as with the visual mixture primaries (it should be noted that the primaries in this usage are not unique in the case of vision; there are a large number of possible mixture primaries); (c) stimuli that are the best exemplars of some sensation, such as NaCl for salty or 585 nanometers for yellow; (d) receptors, neurons, or neural mechanisms that are especially sensitive to a particular stimulus. This last usage often implies a one-to-one correspondence between two or more of the meanings of the term *primary,* for example, primary stimulus – primary neural element – primary sensation. It will be recognized that this sort of isomorphism does not hold in the case of vision. The mixture primaries need not be psychologically primary; the receptor primaries are not the same as the sensation primaries, and the neural primaries are still different. It is little wonder that workers in vision tend to eschew the use of the term *primary* altogether.

The related problems of primaries and psychophysical linking hypotheses are illustrated by Frey's (cited in Melzack & Wall, 1962)

theory of cutaneous sensory coding. As Melzack and Wall pointed out, Frey's theory involves a number of assumptions: (a) that there is a one-to-one correspondence between dimensions of sensory experience and types of receivers in the brain; (b) that there are distinct afferent nerves; (c) that there are distinct receptors in the skin; and (d) that there are distinct stimulus dimensions that subserve the dimensions of sensory experience. The same assumptions are often made in discussions of taste primaries. However, the naive notion of a direct correspondence between levels of processing need not hold for the notion of psychological primaries (the first usage of the term mentioned earlier) to be valid. We prefer the term *basic* taste to *primary* taste because it is used unambiguously to refer to the sensation.

Problems of Gustatory Sensory Coding

There are at least three separate problems to solve before it can be said that we have achieved a theory of sensory coding in taste: (a) the nature of the sensations, (b) the nature of the stimulus dimension(s), and (c) the physiological mechanisms that subserve the coding of the sensations. Each of these is discussed below. As one will see, the analysis of taste sensation has had a longer historical development than that of the other two problems. Our knowledge of the physiological mechanisms involved in the coding process and of the nature of the stimulus dimensions is far from complete. Therefore, one is forced to start with the sensory analysis and work backward to the stimulus to achieve a theory of sensory coding. This process involves what Boynton and Onley (1962) called a *converse linking hypothesis,* that is, inferring stimulus or neural events from sensation rather than the reverse. As Boynton and Onley pointed out, it is much more tenuous to conclude that a particular converse linking hypothesis is true, because many different stimuli can give rise to the same response, whereas the same stimulus (or the same neural activity) should always give rise to the same sensation.

Much of the psychophysical and physiological taste research has involved the converse linking hypothesis either implicitly or explicitly. Thus, the position taken with respect to the nature of taste sensations has an important practical impact on the approach to the investigation of stimulus dimensions and physiological mechanisms. For example, the systematic search for similarities among a large group of stimuli that all give rise to the same taste quality rests on the assumption that there are a limited number (e.g., four) of discrete taste qualities. Birch (1976), for example, reviewed the correlation between the molecular structure of sugars and the intensity of their sweet taste. And similarly, Dastoli and Price (1966) extracted "sweet-sensitive" proteins from the bovine tongue. The choice of exemplars of salty, sour, sweet, and bitter stimuli in almost all psychophysical and physiological research usually rests on the assumpution that these four qualities *are* the taste sensations. On the other hand, by assuming the continuous nature of taste sensations, Schiffman and Erickson (1971) found that taste stimuli formed a multidimensional space that they believed to be incompatible with four basic qualities.

Historical Development of Taste Categories

The origins of the current views on the nature of taste qualities may be found in the historical development of the psychological sensation categories and their relationship to the taste system as it was known prior to electrophysiological recording. Our summary of some of the historical aspects of the taste system is based largely on Bartoshuk's (1978) review of the history of taste.

The naming and categorization of taste sensations date back to the Greeks. Aristotle (384–322 B.C.), for example, listed seven basic tastes: sweet, bitter, sour, salty, astringent, pungent, and harsh. Between the time of the Greeks and the 19th century, the number of taste sensations described as *basic* varied among the seven named by Aristotle, the eleven by Haller (in 1786), an unlimited number by Rudolphi (in 1823), the six by Wundt (in 1880), the five by Öhrwall (in 1891), and the four by Kiesow (in 1896). During the 1800s the investigation of the

anatomy of the taste system prompted the elimination of some previously named basic tastes as actually being the result of olfaction or touch (Bartoshuk, 1978; Öhrwall, 1891; Skramlik, 1926). Thus, by the early 1900s salty, sour, sweet, and bitter were generally agreed upon as basic tastes with alkaline or insipid often also included.

Classical Bases for the Four Taste Qualities

The basic taste qualities were arrived at largely by introspective analysis, but other evidence was used to support their existence. One justification for the four qualities considered was the adaptiveness of the association between the various taste qualities and the nutritive or poisonous characteristics of chemical substances. For example, salt for some animals is a vitally important and scarce nutrient; for others the problem may be to avoid the osmotic consequences of high salinity (Dethier, 1977). Sweet is correlated with high calorie substances, bitter with poisons, and sour with corrosive substances. In fact, analysis of taste stimuli suggests broad and reasonably powerful predictive validity for the taste qualities, particularly between sourness and the H^+ ion, but also between sweetness and carbohydrates, bitterness and alkaloids, and saltiness and alkali halide salts. Other evidence for four qualities was the work showing the differential sensitivity to the four qualities across the tongue (cf. Boring, 1942; Hanig, 1901); the selective disappearance of quality sensitivity following the application of topical anesthetics (Bartoshuk, 1978; Kiesow, 1894b); the selective modification of taste quality following gustatory contact with certain plant substances—*Gymnema sylvestre* and *Synsepalum dulcificum* (Kiesow, 1894b; Shore, 1892; Skramlik, 1926); and the lack of synthetic effects of taste mixtures (Kiesow, 1894a, 1896).

Modality Versus Quality

By the turn of the century the number of the basic tastes was generally agreed upon, but agreement did not exist as to the nature of these taste categories. The views of Helmholtz (1879/1968) on perception in general and his distinction between sensory qualities and sensory modalities in particular had a great influence on those who were to begin the debate on the nature of basic tastes nearly a quarter of a century later. Helmholtz classified sensations as belonging to separate *modalities* if there were no transitions between them; for example, "blue, sweet, warm, and the pitch of tones" (p. 210) belong to separate modalities. The term *quality* was applied to different perceptions within the same sense; for example, blue and violet would be two qualities within color vision.

The modality–quality distinction was applied to taste, and the debate began. On one side were those who thought of salty, sour, sweet, and bitter as four qualities within one sense; the alternative view was that the four basic tastes actually represent four separate sensory modalities. Kiesow, a student of Wundt, was one of many who thought of the four basic tastes as qualities. He went further, however, and proposed a theory to relate these qualities that was based on an analogy to the opponent-process theory of color vision proposed by Hering, with whom he was personally acquainted (Murchison, 1930). Hering's theory was based on his observations of the simultaneous contrast exhibited by the primary visual hues of red and yellow or blue and green. It was Kiesow's view that the qualities salty, sour, sweet, and bitter are actually the four taste "primaries." However, he had only moderate success in demonstrating simultaneous and successive contrast in taste (Kiesow, 1894a, 1896). Öhrwall (1891, 1901), on the other hand, argued that there was no "transition" to be found among any of the four basic tastes and that they were, therefore, four separate modalities. Contrary to Kiesow, he maintained that the phenomena of simultaneous contrast and successive contrast did not exist in taste. Öhrwall criticized Kiesow's taste theory analogy with color vision as faulty, but Kiesow's view that the four tastes are primary qualities prevailed.

Research on the response of the taste system to a mixture of two or more sapid sub-

stances does not support these analogies to color vision. First, no new qualities are produced by a taste mixture, which strongly suggests that taste sensations are limited (Bartoshuk, 1975). That is, a mixture of NaCl and HCl will taste salty and sour but never sweet and/or bitter and/or something else. There is no gustatory analogue to orange. Second, it is not possible to produce a tasteless mixture by an appropriate combination of components, unless the components are themselves so weak as to be nearly tasteless. Thus, there is no gustatory analogue to gray. On these bases, taste has been considered an analytic sense rather than a synthetic sense.

Most researchers preferred to think of salty, sour, sweet, and bitter as four qualities as opposed to the stronger view that they are four separate modalities. However, there was still disagreement among those supporting the more popular *quality* view as to the relationship among those qualities. Applying the label *primary* to each quality and making analogies with theories of color vision dominated and continue to dominate most theories of taste. For example, in the early part of the 20th century, Henning (1927) suggested that the relationships between the four basic tastes could be represented by a hollow tetrahedron. By placing salty, sour, sweet, and bitter on the corners, all other tastes could be placed either along the continuum between two tastes (e.g., on the salty–bitter edge) or on the surface formed by three basic tastes. Thus, all tastes could be represented as appropriate mixtures of two or three of the four "primaries." Representing a much less popular alternative to the primary qualities – color vision analogy, Hahn (1934; Hahn, Kuckulies, & Bissar, 1940; Hahn, Kuckulies, & Taeger, 1938; Hahn & Ulbrich, 1948) argued that salty, sour, sweet, and bitter (and alkaline) are independent qualities (Bartoshuk, 1978). Hahn based his views on his work in adaptation and cross-adaptation and their effects on taste thresholds.

Although there was disagreement on whether the taste system operates like color vision, descriptions of the relationships among different tastes rested on some variation of Müller's influential "doctrine of specific nerve energies" (cited in Boring, 1950). The general assumption was that there is a simple, if unspecified, correspondence between some physical attribute of the stimulus, the operation of the nerve, and the sensation produced. With the development of techniques for recording the electrical activity of single nerve fibers, the search for neural correlates of the human taste experience was broadened to other mammalian nervous systems. Although the historical psychophysical data suggested a search for specific salty, sour, sweet, and bitter fibers, Pfaffmann's (1941) finding that the four basic tastes were not mediated by specific fibers created a problem that is still with us today.

Modern Search for Taste Qualities

With the advent of the vacuum tube and the subsequent development of gustatory electrophysiology, it became apparent that no neat correspondence exists between taste qualities and individual taste nerve fibers (Pfaffmann, 1941). One of the results of this failure of the straightforward extension of the doctrine of specific nerve energies has been to intensify the sort of theorizing typified by Kiesow (1894a, 1894b, 1896), which makes use of the analogy of the senses. It is inevitable that such comparisons be made, and, of course, they are beneficial to the extent that the various senses display common mechanisms of neural functioning. As one sees below, the analogy of the senses has been used to raise a number of objections to the four taste qualities.

Cultural Relativity Objection

The first objection is the notion that the taste qualities are culture bound; that is, if one were to study other cultures, or to throw away the usual categories and start from scratch, one might end up with a completely different set of taste categories. Perhaps. But the best example we have of such a possibility is the number of names for snow em-

ployed by Eskimos. However, this example is not very pertinent because, with some training, any person could discriminate the bases for the distinctions among the names. Studies of non-Western cultures (Chamberlain, 1903; Myers, 1904) find that various languages often lack specific names for one or more of the four taste qualities. Even though one or more of the four may be lacking, one does not seem to find cultures that make more distinctions than the four basic tastes that cannot be explained as describing flavor as opposed to taste, for example, astringent. Recently it has been found that Malay speakers use the same taste words as English speakers but tend to use more modifying adjectives (O'Mahony & Muhiudeen, 1977). There is, of course, a parallel controversy in color vision (e.g., Bornstein, 1973). Heider (1972) studied recognition memory and learning of associations to focal and nonfocal colors by New Guinea Dani, who lack names for hues in their language. (Focal colors are those that best exemplify common color names, e.g., red, orange, yellow, etc.) She found that the Dani made fewer errors in recognition of focal compared to nonfocal colors and learned associations to focal colors faster than to nonfocal colors. Heider also found that native speakers of all language families used shorter words or phrases to describe focal colors than to describe nonfocal colors. Heider's experiments show that the effect of language on perceptual categorization has been overstated, at least for hues. In the absence of similar work in taste, Heider's data imply that taste names are unlikely to be influenced greatly by language. The weight of evidence favors the notion that the lack of comparable names for certain colors in various cultures has a basis in the relative lack of utility of these names in primitive cultures (Woodworth, 1910) rather than in a fundamentally different way of perceiving.

Multidimensional Space Objection

Similar to the cultural relativity objection is the suggestion made by Frings (1948) that "the primary modalities of taste might . . . be simply points of familiarity in an unbroken series of stimulative values. . . . The work of Pfaffmann on electrical impulses from tastebuds, although not fully explicable by this hypothesis, is certainly more nearly explained by it than by the hypothesis of primary modalities" (p. 32). This suggestion makes explicit the notion of a continuum, although none has been identified. Erickson and his co-workers have followed in the direction suggested by Frings (cf. Doetsch, Ganchrow, Nelson, & Erickson, 1969; Erickson, 1963, 1967, 1968; Erickson, Doetsch, & Marshall, 1965; Erickson & Schiffman, 1975). What Erickson and his co-workers were looking for was a taste "wavelength" that would permit an isomorphic correspondence among the chemical, neural, and psychological processes. "Stimuli which are proximate in a multidimensional space on the basis of the similarity of the neural inputs [in terms of across-fiber patterns] are also similar in terms of psychophysical judgments" (Schiffman & Erickson, 1971, p. 632). Erickson et al. (1965) attempted to find an isomorphism between the neural and the chemical domains. They asked, "Is it possible to derive the NRFs [neural response function, i.e., "some measure of the neural activity as a function of a stimulus dimension" (p. 262)] for taste without knowledge of the relevant stimulus dimensions? Further, is it possible to discover the stimulus dimensions? Data presented . . . show that the NRFs and stimulus dimensions for taste may be determined" (pp. 248–249). But even though it might be argued for vision that a sort of isomorphism exists among the stimulus, receptor, neural, and psychological levels, it is a rather rubbery isomorphism that involves, among other things, a shift from three types of receptors to two types of opponent-process neural elements. It is true that the spectrum is preserved through all of these levels but not isomorphically with respect to the number of mechanisms involved. It seems better in the case of vision to postulate something other than a strict isomorphism as a linking hypothesis.

Schiffman and Erickson (1971) have proposed a model for gustatory quality that is

based on multidimensional scaling of taste stimuli. They presented a large number of solutions to subjects who scaled them for similarity and on a number of semantic dimensions. They derived a taste space from this procedure that, they argued, does not support the notion of taste primaries. There are several problems in accepting this conclusion, problems with the experimental method, problems with the multidimensional scaling, and problems in interpretation.

First, and fundamental to the whole argument of their article, they did not take adequate precautions to assure that they stimulated only the sense of taste, as it is usually defined for psychological purposes, namely, taste bud stimulation. Their subjects wore nose plugs "to reduce olfactory input," but this is not sufficient to eliminate the sense of smell. A better procedure would have been to use the technique of forcing a gentle stream of air into the nostrils (Mozell, Smith, Smith, Sullivan, & Swender, 1969), thereby preventing the reflux of odorous air into the nasal cavity. This is a curious procedural deficiency, since the purpose of the experiment was to determine the dimensionality of a taste space. Any contribution from the sense of smell seriously distorts the taste space. This weakness is particularly significant considering that the history of taste qualities shows a trend toward a decrease in the number of taste qualities as a result of increasing care in eliminating sensations mediated by other sensory systems. It was such careful psychophysical work that ruled out the alkaline and metallic tastes many years ago. Skramlik (1926) reviewed work by Frey and Herlitzka that showed clearly that the alkaline and metallic tastes resulted from olfactory sensations and disappeared when these were excluded. Another standard method of determining whether a sensation quality can be attributed to taste is to place the substance on parts of the tongue, such as the center, that are known to be devoid of taste buds. Such observations may well have obviated discussion of other nontaste qualities like *bitey, burning,* and *tingling.* Skramlik's review demonstrates clearly that the earlier workers were keenly aware of these problems in the psychological analysis of taste qualities.

The scaling problem in the Schiffman and Erickson (1971) experiment is inherent in the use of multidimensional scaling. Although multidimensional scaling has been used a great deal with considerable impact in certain areas, it may fairly be said that its contribution to sensory processes has not been the discovery of unknown dimensions as much as the improvement of our understanding of the metric relationships among known dimensions. Further, multidimensional scaling is extremely sensitive to the choice of stimuli to be scaled. If most of the stimuli are very similar to one another, one sort of dimensionality will result. If a single stimulus is added that is very different from all of the other stimuli in the sample, an entirely new dimension will emerge as a result. For this reason, scaling of a small subset of stimuli often gives fundamentally different results from the scaling of the entire set. This is particularly true in smell (e.g., Schiffman, 1974) and taste (e.g., Gregson, 1966). In addition, different multidimensional scaling methods often give different results. This can be seen in the data of Schiffman and Erickson (1971), in which the scaling based on similarity led to three dimensions and the scaling based on the semantic differential data gave two. Schiffman and Erickson chose to accept the similarity data for the space, but they interpreted the dimensions by means of semantic differential scales. On this basis they came up with the following three dimensions: molecular weight, hedonic, and deviation of pH from neutrality. Keeping in mind that the method of stimulus presentation and the scaling technique are both weak, the interpretation of the space presents problems. First, one of the dimensions, hedonic, is a purely psychological dimension. The other two, pH and molecular weight, are purely physical. Second, these dimensions are not orthogonal; hedonic is correlated with molecular weight. In addition, there are some large inconsistencies in the locations of stimuli along the molecular weight dimension. It is difficult to see how such a space could be useful in devising linking hy-

potheses about the coding of taste qualities. In any case, it is fair to say that the four taste qualities fall into four clusters in their space. It appears that the data would be at least as well described by the four traditional taste qualities, with those that fall outside of the space defined by the four qualities representing, in part, extragustatory (e.g., smell) sensation. What Schiffman and Erickson have achieved is a description of a psychological taste space that neither supports nor denies the existence of four basic tastes.

To return to the problem of interpretation of the multidimensional scaling data, consider the multidimensional scaling of colors (cf. Schiffman & Erickson, 1971), which to date has accomplished the validation of the Munsell color space. Although this is an achievement in its own right, it in no way constitutes a theory of sensory coding, in and of itself, because it is simply a description of the psychophysical relationships among the colors in terms of the physical stimulus.

Multidimensional scaling of stimuli presented to different senses has been performed. For example, Wicker (1968) had subjects rate the similarities between pairs of tones, Munsell color chips, or tones and chips in an effort to study synesthetic relationships among these two stimulus domains. He found that a single multidimensional space accounted well for both the colors and the tones. Other intersensory interactions are well-known, as the literature on auditory–visual synesthesia demonstrates (e.g., Marks, 1975). Although not all people experience synesthesia, and it may seem somewhat artificial to some, nonsynesthetes relate color dimensions to auditory dimensions in the same way as synesthetes; that is, subjects are very willing to relate sensations from systems that are undeniably distinct. Nafe (1927), using introspection, analyzed the qualities of the sensations mediated by the skin senses. He concluded "that the 'qualities' of felt experience . . . are analyzable patterns of experience . . . that such experiences vary in brightness, volume, density, and definiteness of outline" (p. 398). Had Nafe been working 50 years later he might have done his experiment using multidimensional scaling. The results he did obtain using the best techniques of his day are consistent with our conclusion that multidimensional scaling does not necessarily yield the underlying sensory dimensions. In fact, multidimensional scaling is often done on very diverse stimuli. The dimensions derived are not taken to represent any physical relationship among the stimuli. For example, in scaling animals, dimensions such as fierceness have been obtained (Henley, 1969). These are not interpreted as physical dimensions or taxonomic classifications (cf. Martindale & Hines, 1975). In addition, various interactions among the various senses are well-known. For example, cold and warm seem to be different sensory systems as far as their anatomy and physiology are concerned, even though their perceptual unity could be argued. Nevertheless, they obviously interact to a great extent, and their interaction may even be responsible for the sensation of heat.

Therefore, little seems to be gained from the multidimensional analysis of taste experience as far as the question of the nature of the taste qualities is concerned. It follows that an isomorphism between the stimulus and the sensation is not the only, or even the best, conclusion to draw from multidimensional scaling of sensation.

Recent Evidence for the Four Qualities

Modern research has made much use of the four taste qualities, but more out of convenience than out of conviction that they are basic. For example, Bartoshuk (1974) wrote, "In light of this [controversy over whether these taste qualities are primaries], some investigators have come to follow a somewhat different strategy in studying taste quality. Instead of concentrating on finding all primaries, they have turned to studying the functional properties of sweet, sour, salty and bitter" (p. 279). Even so, evidence for the limited number of taste categories comes from recent psychophysical studies of the effects of the following: cross-adaptation, water taste, drugs, taste-altering substances (*Gymnema sylvestre* and miracle fruit), mixtures, and spatial and temporal properties. It is true that taken alone these data have a

certain circularity about them as evidence for the four basic tastes. The four taste qualities were used as response categories and may be seen as forcing the data in the direction of evidence for basic tastes. It should be noted, however, that subjects in some of the experiments discussed below were given the opportunity to use additional categories to describe their taste sensations, but did not do so with any regularity (e.g., Bartoshuk, McBurney, & Pfaffmann, 1964). The point we wish to emphasize is that the four taste qualities are both necessary and sufficient to account for the qualitative and quantitative differences in the psychophysical data.

Cross-Adaptation

Psychophysical studies of the response of the taste system to one stimulus following adaptation to another support the limited number of taste sensations. It is true that this work on cross-adaptation could not have been done without having the subjects report on the taste qualities of the compounds. But it is clear that the effect of adaptation is to eliminate or substantially reduce the response to the quality to which the tongue has been adapted, whatever its name. Where the cross-adaptation was between stimuli sharing the same quality, the following results were reported: (a) Sucrose adaptation reliably reduced the sweetness of all sweet-tasting compounds tested (McBurney, 1972); (b) adaptation to NaCl reduced the saltiness of every salt tested (Smith & McBurney, 1969); (c) adaptation to citric acid reliably reduced the sourness of all other acids tested (McBurney, Smith, & Shick, 1972); and (d) the results for bitter were not as clear-cut, since adaptation to quinine HCl reduced the bitterness of some bitter compounds without affecting others (McBurney et al., 1972).

The psychophysical data on cross-adaptation have implications for the number of receptor mechanisms that must operate in the taste system. The effects of cross-adaptation are found to a greater or lesser extent in all four qualities, and there is little across-quality adaptation, thus implying the operation of separate receptor mechanisms. The data do not prove the existence of specific receptor sites, but they do imply that there are a limited number of receptor mechanisms. For example, it is likely that a single receptor mechanism codes the sweet taste, another codes saltiness, and a third codes sourness. The mechanism for bitterness, however, is more complex.

Water Taste

The water taste phenomenon, that is, the taste of water after adaptation to a compound, also supports the limited number of taste sensations. Each of the four qualities and no others have been produced as water tastes after suitable adaptation (Bartoshuk, 1968; McBurney, 1969; McBurney & Bartoshuk, 1972; McBurney & Shick, 1971). It should be noted that the water taste phenomenon cannot be explained by a simple opponent process. For example, adaptation to NaCl will cause water to have a bitter or sour–bitter taste; bitter and sour substances induce a sweet water taste; sweet substances produce a sour or bitter water taste; and salty water tastes have only been reliably produced by urea and closely related compounds. Apparent cross-enhancement between qualities was found in some studies that used the cross-adaptation paradigm. However, the increase in the perceived intensity of a stimulus after adaptation was shown to be explained by the water taste phenomenon (McBurney & Bartoshuk, 1973). The increase was always in the quality normally produced as a water taste rather than in the dominant quality of the compounds.

Taste-Altering Substances

The specificity of the effects of two taste-altering substances, found in *Gymnema sylvestre* and *Synsepalum dulcificum* (miracle fruit), has provided information about the nature of taste sensations and possible taste receptor mechanisms. The replication and extension of some of the historical work on the effects of chewing *Gymnema sylvestre* showed that only the sensation of sweet is abolished, including the sweet water taste

(Bartoshuk, Dateo, Vandenbelt, Buttrick, & Long, 1969). Since stimuli that elicit a single taste quality (sweet) are all influenced in the same way by *Gymnema sylvestre,* the implication is that the taste quality of sweetness is distinct from the other qualities. There is also the implication that there is a single receptor mechanism that codes the sweet taste and that the gymnemic acid operates by competing with sweet compounds for the sweet receptor sites. Miracle fruit is a berry that, when chewed, causes sour substances to taste sweet. The effect of this taste modifier can be explained by analyzing its effect on the particular qualities of sour and sweet. K. Kurihara (1971) and K. Kurihara, Y. Kurihara, and Beidler (1969) suggested that the action of miracle fruit is a result of the interaction on the taste cell membrane between the acid of a normally sour substance and the glycoprotein that is the active ingredient in the fruit. They suggested that the glycoprotein does not itself normally stimulate any receptor sites but that the presence of acid causes a change in the shape of the membrane in such a way as to cause contact between the glycoprotein and sweet sites. Thus, it is possible to understand the sweetening effect as the addition of a new taste (sweet) and the lack of increase in total intensity as a case of mixture suppression of sour by sweet (Bartoshuk, 1975).

Effect of Locus

A reexamination of the variation in sensitivity over the tongue has produced evidence for quality specificity and some insight into the complexity of the receptor mechanisms that operate in the taste system. Collings' (1974) investigation of the effect of locus of stimulation on taste threshold generally confirms Hanig's (1901) data on the differential sensitivity of the tongue and supports the distinctiveness of the taste sensations. The exception she found was that sensitivity to bitter was greater at the front of the tongue than at any other tongue locus, not at the back as Hanig reported.

Stevens (1969) argued that the exponent of the power function that relates sensation magnitude to stimulus concentration is principally determined by the characteristics of the receptor system. (Even though fairly consistent differences are found among the exponents for the four qualities [Meiselman, 1972], we have not used this as evidence for basic tastes. The exponent is sensitive to several variables in addition to quality, namely, method of stimulation; range of stimuli; psychophysical method; presence, value, and location of standard; and so on [Stevens, 1975]. Furthermore, the exponents have never been determined with the exactness that would justify making a strong case on this basis.) Evidence for the distinctness of receptor mechanisms for taste qualities comes from the fact that, in general, locus of stimulation also has a differential effect on the psychophysical function that relates sensation to stimulus concentration, both between qualities and within one quality (Collings, 1974). The differences Collings found between qualities are compatible with the existence of four distinct sensations, but the differences she found across loci within qualities imply a complex receptor mechanism that is difficult to explain. Because Smith (1971) showed that varying the number of receptors stimulated (at the front of the tongue) does not change the psychophysical function for a given taste compound, Collings concluded that the observed differences in threshold were not simply the result of differences in the number of particular receptors present at each location.

Temporal Properties

Another aspect of taste that has been studied is the temporal properties of the gustatory system (cf. McBurney, 1976). Early research (1914–1955) on the temporal response of the taste system falls into four areas: reaction time (Bujas, 1935; Piéron, 1914), growth of sensation (Bujas & Ostojcic, 1939), relationship between time and intensity (Bujas, 1934; Hara, 1955), and rate of adaptation (Bujas, 1953; Hahn, 1949). The first three areas are closely related, and the early research indicates that

the temporal response of the system varies according to the quality (e.g., salty, sour, sweet, or bitter) of the stimulus. In particular, reaction times to a stimulus were found to be a hyperbolic function of intensity and varied among qualities in increasing order from salty, sour, and sweet to bitter. A similar ordering was found to hold for growth of sensation. Results concerning rate of adaptation are not so clear. Problems with the methods used to measure rate of adaptation and questions concerning the existence of complete adaptation prompted the search for another technique for studying temporal properties.

The most recent psychophysical method of investigating the nature of taste qualities is the application of the techniques of linear systems analysis to the gustatory system. Work such as this has been going on for some time in vision (e.g., Sekuler, 1974). In our laboratory we attempted to investigate the temporal properties of the taste system (McBurney, 1976). Briefly, linear systems techniques can be used to study the input–output relationships of a linear system. By finding the ratio of a sinusoidally varying input to the output across a range of frequencies, a transfer function can be calculated that is characteristic of the temporal properties of the system. Extension of this technique to the gustatory system was made by presenting the tongue with a solution of sinusoidally varying concentration. The output studied was the threshold for several frequencies of stimulus presentation. A temporal-modulation sensitivity function was obtained for each of the four qualities. Our results agree with the earlier data in that the ordering of the sensitivities of the four qualities as represented by the sensitivity functions was as follows: salty, sweet, sour, and bitter. These functions differentiated among qualities and therefore imply that the functions represent temporal characteristics of four separate systems.

Inner Psychophysics of Taste

Because the approach of this article has been psychophysical, it is appropriate here to point out a distinction made over 100 years ago by the founder of psychophysics, G. T. Fechner—that between outer and inner psychophysics (Howes & Boring, 1966). Outer psychophysics is the familiar enterprise of relating sensation to external stimulus. Inner psychophysics is the relationship of sensation to the physiology of the nervous system. Of course in Fechner's day this was only a far-off dream, but we have made some progress along these lines in recent years. The purpose of this section is to review the inner psychophysics of taste, the relationship of the physiological data on neural mechanisms (receptor sites, cells, and fibers) to the existence of taste qualities. We discuss the implications these data may have for linking hypotheses concerning the neural coding of taste quality.

Receptor Sites

The nature of the receptor site(s) for taste is still poorly understood. However, Beidler (1971) has determined that NaCl and sucrose each bind to one receptor site and that KCl, NH_4Cl, and other salts that do not have pure salty tastes bind to more than one site. Such a finding is consistent with the position that the four qualities are basic tastes. If the taste of NaCl were but one of many tastes located along a single taste dimension, then one might expect every substance to stimulate a separate site or expect NaCl or sucrose, for example, to stimulate more than one.

Neural Coding of Quality

Pfaffmann, in his classic study (1941), and all subsequent investigators that used vertebrate subjects found that single taste neurons typically respond to stimuli representative of more than one taste quality. As a result, Pfaffmann (1959a, 1959b) proposed that the sensory code for taste quality is carried by the relative firing rate of two or more fibers. This hypothesis linking the nonspecific neurons to specific sensations has been called the across-fiber pattern theory. Zotterman (1959), on the other hand, emphasized that each fiber typically has a stim-

ulus to which it responds best and argued that the activity in each fiber signals the quality of its best stimulus. This position has come to be known as the labelled line theory. The difference between the two hypotheses has not yet been resolved, since the weight of experimental evidence does not clearly favor the across-fiber pattern or the labelled line.

For example, Frank (1973) recorded from the chorda tympani of the hamster and found the same sort of broad tuning that was found by Pfaffmann in the rat. However, she discovered that if the stimuli were arranged in the order sweet, salty, sour, bitter, there was always one maximum of responding, with responsiveness decreasing monotonically on either side of the maximum. Thus, instead of the responsiveness of the neurons to various qualities that appear to be random, there was some orderliness for the first time since the beginning of electrophysiological study of taste neurons. Frank pointed out that the order of stimuli that produces the simplification of the profiles happens to be from acceptance to rejection. This is a curious finding and may simply be fortuitous. However, Frank's results have led to a renewed interest in the idea that the sensory code may be contained in a single neuron, or in the labelled line, as it has come to be called.

Comparison of Psychological and Psychophysical Data Relative to Coding

Direct comparisons of neural and psychophysical responses in the human are obviously limited to special circumstances, but such studies have been done. Electrophysiological responses of the exposed chorda tympani nerve of otosclerotic patients were found to be closely correlated with their magnitude estimates of the intensity of various sweet stimuli. Also, the neural as well as the psychophysical response to sweet stimuli was abolished following the application of *Gymnema sylvestre*. In addition, it was found that the time required for complete adaptation to NaCl agreed with the psychophysical reports (Borg, Diamant, Oakley, Strom, & Zotterman, 1967; Diamant, Oakley, Strom, Wells, & Zotterman, 1965).

The electrophysiological investigation of the effects of cross-adaptation has produced data that are remarkably similar to those from the psychophysical studies mentioned above. Smith and Frank (1972) studied cross-adaptation betwen salts in the rat chorda tympani and found that the transient response to a salt was affected by adaptation to other salts. They compared the amount of cross-adaptation they obtained with that obtained by Smith and McBurney (1969) for those pairs of stimuli that were common to the two experiments. The degree of cross-adaptation observed was very similar in the two studies. In addition, it was found in both studies that salts with common cations cross adapted more than did salts with different cations. The implication of these two studies is the same, namely, that there is a common neural mechanism for saltiness. As discussed earlier, this does not necessarily imply that separate receptor sites or fiber types exist. In fact, the data of Smith and Frank (1972) are very similar to the correlations obtained by Erickson et al. (1965) from the responses of individual rat chorda tympani fibers to pairs of salts and are thus equally compatible with an across-fiber pattern theory of quality coding.

The most determined attempt to extend Pfaffmann's across-fiber pattern theory has been made by Erickson. From the relative responsiveness of each neuron to pairs of salts, Erickson et al. (1965) derived what they termed a *neural response function* (NRF) for each neuron that described its sensitivity to many salts and that allowed them to order the salts as if they fell along a continuous dimension. They reasoned that the NRFs for gustatory neurons were similar to the relative activity of visual neurons in response to the visual spectrum. The crucial assumption of Erickson's approach to taste coding is the existence of a continuous stimulus dimension, and all of Erickson's hypotheses about sensory coding are based on this assumption. It is because of this assumption that the NRF work suffers from the same weakness as the multidimensional scaling work described

earlier. Specifically, it relies on a converse linking hypothesis, in this case a converse physiological–physical linking hypothesis. Erickson attempted to find the stimulus dimension from an analysis of the relationships among the physiological responses. Such an approach will lead to a solution, but other interpretations of the results of the NRF work are possible. For example, one could do the physiological equivalent of the multidimensional scaling of cutaneous stimuli discussed above as a thought experiment. This time consider establishing NRFs for many stimuli presented to the chorda tympani nerve, which subserves touch and temperature as well as taste. It is likely that NRFs could be established for these stimuli and that one could develop some sort of space based on them. However, it would not demonstrate the existence of a physical stimulus dimension that underlies all of the various stimuli.

The Analytic–Synthetic Distinction and Coding Theory

Erickson's work was a very creative extension of Pfaffmann's approach to taste coding, and one of its most important contributions was to make explicit a number of hypotheses about sensory coding. However, it is clear that there are difficulties created by the assumption of a single stimulus dimension, as, for example, in Erickson's explicit hypothesis about the coding of taste mixtures. Erickson (1968) claimed that synthetic systems are coded by across-fiber patterns and analytic systems by specific fibers. On this basis, taste must be a synthetic system in spite of the fact that the qualities of the components in a mixture are not fused or otherwise lost and that no new qualities are produced. It is true that color vision is both synthetic and has a pattern theory of coding. But, audition is clearly analytic, and the auditory nerve displays fairly broad tuning at the level of the first-order neuron (Kiang & Moxon, 1974). Relatively specific firing is found only at more central levels. Therefore, the correlation Erickson proposed does not seem to hold. Erickson suggested that the mixture of two stimuli in a sense that requires a pattern theory will result in a new pattern that has a different maximum than the two original patterns and hence in a new (synthetic) sensation. One can, in fact, match a stimulus like KCl, which is salty and bitter, with a mixture of NaCl and quinine, and a subject may not be able to distinguish such a mixture from KCl. It should be noted, however, that KCl tastes salty and bitter and not something else. The "missing orange" seems a strong argument against taste being synthetic. But, the question of whether KCl is itself a mixture to begin with is actually begged. Erickson has suggested that KCl may be a mixture taste: The taste of NaCl itself could be considered a mixture taste because the stimulus solution is composed of two ionic components (Na^+ and Cl^-) that might be located at two different points along the stimulus dimension (Erickson & Schiffman, 1975). This suggetion cannot be dealt with here in detail except to say that Beidler's (1965, 1970, 1971) theory of sensory transduction in taste, which is the most widely accepted theory, considers the cation to be responsible for stimulation and the anion to be inhibitory. Further, sour, sweet, and bitter tastes cannot be similarly considered mixture tastes in this context because the stimuli are either nonionic (sweet and bitter) or the stimulation is accepted to be due to the cation (acids). This seems to create a problem for the assumption of a continuous stimulus dimension.

Two tests of the analytic–synthetic question have recently been reported. Erickson (1977) tested reaction times of subjects who had to choose between high and low concentrations of NaCl when $MgCl_2$ also varied in concentration and vice versa. He found that variation in the second compound had an effect on reaction time, being facilitating when the two covaried and inhibiting when they varied independently. However, other interpretations are possible. Total intensity of the mixture would be a relevant cue in the covarying condition and would be a distractor in the independently varying condition. Further, ignoring the overall intensity of the mixture, one has two qualities that vary independently in intensity. There are a

number of studies that suggest that an analytic sense would not necessarily code these two qualities independently. Hamlin, Stone, and Moskowitz (1955) had subjects sort cards according to shape of symbol. Performance was slower when the symbols varied in color than when they were always black. Egeth (1967) and Well (1971) reviewed other similar experiments. Although performance has not always been found to suffer in such experiments, the case made by such a design for the synthetic nature of taste is extremely tenuous. Nowlis and Frank (1977) tested the synthetic–analytic question by using the conditioned aversion paradigm (Garcia, Hankins, & Rusiniak, 1974) with rats. Rats that have been poisoned after drinking a solution will avoid that solution in a later test. If they are poisoned after drinking two solutions, one of which is familiar and one novel, they will avoid only the novel solution. In a very clever design, Nowlis and Frank made their rats sick on a mixture of NaCl and sucrose. When the sucrose was the familiar stimulus, they later avoided the NaCl as much as they avoided the mixture and did not avoid the sucrose. Results were similar when NaCl was the familiar stimulus. When neither was familiar the rats avoided both sucrose and NaCl. This is strong evidence that the rats sorted out the two components of the mixture as separate and independent tastes. Thus, Erickson's (1977) hypothesis linking across-fiber patterning to synthetic systems does not seem to hold. The evidence argues against taste being a synthetic sense.[1]

Summary

In summary, support for basic tastes in the gustatory system is found in psychophysical work, both historical and modern. First, the history of thinking on taste has shown a narrowing of the number of taste qualities to the present four on the basis of careful introspective work. Classical support for the four taste qualities also included the association between taste quality and the nutritive or poisonous nature of the chemical, the correlation between chemical structure and taste, the differential sensitivity of the four qualities across the tongue, the effects of topical anesthetics, and the lack of synthetic effects of mixtures. Recent support for the four taste qualities comes from work on cross-adaptation, water taste, taste-altering substances, the effects of locus on the threshold and on the growth of sensation with stimulus intensity, and the temporal properties of the taste system.

Electrophysiological work on taste coding began by looking for nerve fibers that would correlate with the four taste qualities. When that appeared to fail, some investigators felt compelled to abandon the concept of the four qualities as basic. However, attempts to infer the dimensionality of the stimulus and solve the sensory code for taste without making use of the four qualities have not been productive. Either of the two major positions on sensory coding in taste, labelled line or pattern theory, appears to us to be equally compatible with the existence of basic tastes. In addition, the question of the analytic or synthetic nature of taste experience seems independent of both problems, namely, basic tastes versus a multidimensional taste space and the labeled line versus the pattern theory.

What Is a Sense?

Implicit in much of the discussion so far has been the idea that taste is a single sense or sensory system. To say that the four qualities exist raises again the question of whether they might well be considered separate sensory systems. Often the four qualities have been called submodalities, implying more distinctiveness among them than is implied by

[1] After this article was written, an article by Dethier (1978) appeared that discusses the question of taste qualities from the comparative point of view. Although he does not reject the four taste qualities for man, Dethier feels that they constitute a straightjacket when applied to other animals. We freely accept that many other animals have different taste worlds. However, as argued in the present article, we believe that the behavioral and physiological work on mammals supports the substantial similarity of taste mechanisms in man and other mammals. Our main point, that taste qualities are not arbitrary conventions, is fully compatible with Dethier's position that there may be either more or fewer than four taste qualities in other organisms.

the primary color sensations but less than is implied by the difference between, say, red and F sharp. Can one imagine a smooth transition between sweet and salty or salty and bitter? We deny this and argue that they meet Helmholtz's (1879/1968) criterion for separate senses. But we must admit that there is room for difference of opinion. However, that the four taste qualities share the same receptors, neurons, and neural projections seems not to be especially important when one considers that the skin senses have a similar degree of commonality among them. We feel it may be more useful to consider taste as comprising four senses: salty, sour, sweet, and bitter. However, this is not to say that the four taste modalities are as distinct as vision is from audition. Clearly vision and audition are different sensory systems because they not only meet Helmholtz's criterion but also because they have separate receptors, neural pathways, and projections in the sensory cortex of the brain. And the skin senses, although together considered to make up the cutaneous system, are separate from each other largely because of the psychological distinctiveness of the various sensations, even though there seem to be separate receptors in some cases.

Thus, the listing of senses must remain somewhat arbitrary. The "special" senses, vision and audition, are relatively but not completely distinct (synesthesia), and the taste senses are related but not unitary.

If one must make analogies among the senses, and one must, then we argue that the proper senses with which one should compare taste are the skin senses. After all, the sense of taste shares the same receptor surface, nerves, neurons, and central projections as do the other skin senses. When one makes these analogies, one sees that the notion of four basic qualities or submodalities of taste has a great deal of empirical evidence to recommend it and avoids the pitfalls into which the analogy with vision has led us.

References

Bartoshuk, L. M. Water taste in man. *Perception & Psychophysics*, 1968, *3*, 69–72.

Bartoshuk, L. M. Taste illusions: Some demonstrations. *Annals of the New York Academy of Sciences*, 1974, *237*, 279–285.

Bartoshuk, L. M. Taste mixtures: Is mixture suppression related to compression? *Physiology & Behavior*, 1975, *14*, 643–649.

Bartoshuk, L. M. History of taste research. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 6A). New York: Academic Press, 1978.

Bartoshuk, L. M., Dateo, G. P., Vandenbelt, D. J., Buttrick, R. L., & Long, L., Jr. Effects of *Gymnema sylvestre* and *Synsepalum dulcificum* on taste in man In C. Pfaffmann (Ed.), *Olfaction and taste III*. New York: Rockefeller University Press, 1969.

Bartoshuk, L. M., McBurney, D. H., & Pfaffmann, C. Taste of sodium chloride after adaptation to sodium chloride: Implications for the "water taste." *Science*, 1964, *143*, 967–968.

Beidler, L. M. Anion influences on taste receptor response. In T. Hayashi (Ed.), *Olfaction and taste II*. Oxford, England: Pergamon Press, 1965.

Beidler, L. M. Physiological properties of mammalian taste receptors. In G. E. Wolstenholme & J. Knight (Eds.), *Symposium on taste and smell in vertebrates*. London: Churchill, 1970.

Beidler, L. M. Taste receptor stimulation with salts and acids. In L. M. Beidler (Ed.), *Handbook of sensory physiology* (Vol. 4, Pt. 2). Berlin, West Germany: Springer-Verlag, 1971.

Birch, G. G. Structural relationships of sugars to taste. *Critical Reviews in Food Science and Nutrition*, 1976, *8*, 57–95.

Borg, G., Diamant, H., Oakley, B., Strom, L., & Zotterman, Y. A comparative study of neural and psychophysical responses to gustatory stimuli. In T. Hayashi (Ed.), *Olfaction and taste II*. Oxford, England: Pergamon Press, 1967.

Boring, E. G. *Sensation and perception in the history of experimental psychology*. New York: Appleton-Century-Crofts, 1942.

Boring, E. G. *A history of experimental psychology* (2nd ed.). New York: Appleton-Century-Crofts, 1950.

Bornstein, M. H. Color vision and color naming: A psychophysical hypothesis of cultural difference. *Psychological Bulletin*, 1973, *80*, 257–285.

Boynton, R. M., & Onley, J. W. A critique of the special status assigned by Brindley to "psychophysical linking hypotheses" of "class A." *Vision Research*, 1962, *2*, 383–390.

Brindley, G. S. *Physiology of the retina and the visual pathway*. Oxford, England: Alden Press, 1960.

Bujas, Z. Le temps d'action des stimuli de la sensibilité gustative. *Comptes Rendue des Séances de la Société de Biologie*, 1934, *116*, 1307–1309.

Bujas, Z. Le temps de réaction aux excitations gustatives d'intensité différente. *Comptes Rendus des Séances de la Société de Biologie*, 1935, *119*, 1360–1362.

Bujas, Z. L'adaptation gustative et son mécanisme. *Acta Instituti Psychologici Universitatis Zagrebensis*, 1953, *17*, 1–11.

Bujas, Z., & Ostojcic, A. L'évolution de la sensation gustative en fonction du temps d'excitation. *Acta Instituti Psychologici Universitatis Zagrebensis,* 1939, *3,* 1–24.

Chamberlain, A. F. Primitive taste-words. *American Journal of Psychology,* 1903, *14,* 146–153.

Collings, V. B. Human taste response as a function of locus of stimulation on the tongue and soft palate. *Perception & Psychophysics,* 1974, *16,* 169–174.

Dastoli, F. R., & Price, S. Sweet-sensitive protein from bovine taste buds: Isolation and assay. *Science,* 1966, *154,* 905–907.

Dethier, V. G. The taste of salt. *American Scientist,* 1977, *65,* 744–750.

Dethier, V. G. Other tastes, other worlds. *Science,* 1978, *201,* 224–228.

Diamant, H., Oakley, B., Strom, L., Wells, C., & Zotterman, Y. A comparison of neural and psychophysical responses to taste stimuli in man. *Acta Physiologica Scandinavica,* 1965, *64,* 67–74.

Doetsch, G. S., Ganchrow, J. J., Nelson, L. M., & Erickson, R. P. Information processing in the taste system of the rat. In C. Pfaffmann (Ed.), *Olfaction and taste III.* New York: Rockefeller University Press, 1969.

Egeth, H. Selective attention. *Psychological Bulletin,* 1967, *67,* 41–57.

Erickson, R. P. Sensory neural patterns and gustation. In Y. Zotterman (Ed.), *Olfaction and taste.* Oxford, England: Pergamon Press, 1963.

Erickson, R. P. Neural coding of taste quality. In M. R. Kare & O. Maller (Eds.), *The chemical senses and nutrition.* Baltimore, Md.: Johns Hopkins University Press, 1967.

Erickson, R. P. Stimulus coding in topographic and nontopographic afferent modalities: On the significance of the activity of individual sensory neurons. *Psychological Review,* 1968, *75,* 447–465.

Erickson, R. P. Role of primaries in taste research. In J. LeMagnen & P. Macleod (Eds.), *Olfaction and taste VI.* London: Information Retrieval, 1977.

Erickson, R. P., Doetsch, G. S., & Marshall, D. A. The gustatory neural response function. *Journal of General Physiology,* 1965, *49,* 247–263.

Erickson, R. P., & Schiffman, S. S. The chemical senses: A systematic approach. In M. S. Gazzaniga & C. Blakemore (Eds.), *Handbook of psychobiology.* New York: Academic Press, 1975.

Frank, M. An analysis of hamster afferent taste nerve response functions. *Journal of General Physiology,* 1973, *61,* 588–618.

Frings, H. A contribution to the comparative physiology of contact chemoreception. *Journal of Comparative and Physiological Psychology,* 1948, *41,* 25–34.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. Behavioral regulation of the milieu interne in man and rat. *Science,* 1974, *185,* 824–831.

Gibson, J. *The senses considered as perceptual systems.* Boston: Houghton Mifflin, 1966.

Gregson, R. A. M. Theoretical and empirical multidimensional scalings of taste mixture matchings. *British Journal of Mathematical and Statistical Psychology,* 1966, *19,* 59–76.

Hahn, H. Die adaptation des geschmackssinnes. *Zeitshrift für Sinnesphysiologie,* 1934, *65,* 105–145.

Hahn, H. *Beitrage zur reizphysiologie.* Heidelberg, West Germany: Scherer, 1949.

Hahn, H., Kuckulies, G., & Bissar, A. Eine systematische untersuchung der geschmacksschwellen II. *Zeitschrift für Sinnesphysiologie,* 1940, *68,* 185–260.

Hahn, H., Kuckulies, G., & Taeger, H. Eine systematische untersuchung der geschmacksschwellen. *Zeitschrift für Sinnesphysiologie,* 1938, *67,* 259–306.

Hahn, H., & Ulbrich, L. Eine systematische untersuchung der geschmacksschwellen. *Pflügers Archiv für die Gesamte Physiologie des Menschen und der Tiere,* 1948, *250,* 357–384.

Hamlin, R. M., Stone, J. T., & Moskowitz, M. J. Rorschach color theories as reflected in simple card sorting tasks. *Journal of Projective Techniques,* 1955, *19,* 410–415.

Hanig, D. P. Zur psychophysik des geschmackssinnes. *Philosophische Studien,* 1901, *17,* 576–623.

Hara, S. Interrelationships among stimulus intensity, stimulated area and reaction time in the human gustatory sensation. *Bulletin of Tokyo Medical and Dental University,* 1955, *2,* 147–158.

Heider, E. R. Universals in color naming and memory. *Journal of Experimental Psychology,* 1972, *93,* 10–20.

Helmholtz, H. von. [The facts of perception.] In R. M. Warren & R. P. Warren (Eds.), *Helmholtz on perception: Its physiology and development.* New York: Wiley, 1968. (Originally published, 1879.)

Henley, N. M. A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior,* 1969, *8,* 176–184.

Henning, H. Psychologische studien am geschmackssinn. In E. Abderhalden (Ed.), *Handbuch der Biologischen Arbeitsmethoden.* Berlin, Germany: Urban & Schwarzenberg, 1927.

Howes, D. H., & Boring, E. G. (Eds.). *Gustav Fechner: Elements of psychophysics.* New York: Holt, Rinehart & Winston, 1966.

Kiang, N. Y-S., & Moxon, E. C. Tails of tuning curves of auditory nerve fibers. *Journal of the Acoustical Society of America,* 1974, *55,* 620–630.

Kiesow, F. Beitrage zur physiologischen psychologie des geschmackssinnes. *Philosophische Studien,* 1894, *10,* 329–368; 523–561. (a)

Kiesow, F. Ueber die wirkung des cocain und der Gymnemasaure auf die schleimhaut der zunge und des mundraums. *Philosophische Studien,* 1894, *9,* 510–527. (b)

Kiesow, F. Beitrage zur physiologischen psychologie des geschmackssinnes. *Philosophische Studien,* 1896, *12,* 255–278.

Kurihara, K. Taste modifiers. In L. M. Beidler (Ed.), *Handbook of sensory physiology* (Vol. 4, Pt. 2). Berlin, West Germany: Springer-Verlag, 1971.

Kurihara, K., Kurihara, Y., & Beidler, L. M. Isolation and mechanism of taste modifiers: Taste-modifying protein and gymnemic acids. In C. Pfaffmann (Ed.), *Olfaction and taste III.* New York: Rockefeller University Press, 1969.

Marks, L. E. On colored-hearing synesthesia: Cross-modal translations of sensory dimensions. *Psychological Bulletin,* 1975, *82,* 303–331.

Martindale, C., & Hines, D. Dimensions of olfactory quality. *Science,* 1975, *188,* 74–75.

McBurney, D. H. Effects of adaptation on human taste function. In C. Pfaffmann (Ed.), *Olfaction and taste III.* New York: Rockefeller University Press, 1969.

McBurney, D. H. Gustatory cross adaptation between sweet-tasting compounds. *Perception & Psychophysics,* 1972, *11,* 225–227.

McBurney, D. H. Are there primary tastes for man? *Chemical Senses and Flavor,* 1974, *1,* 17–28.

McBurney, D. H. Temporal properties of the human taste system. *Sensory Processes,* 1976, *1,* 150–162.

McBurney, D. H., & Bartoshuk, L. M. Water taste in mammals. In D. Schneider (Ed.), *Olfaction and taste IV.* Stuttgart, West Germany: Wissenschaftliche Verlagsgesellschaft, 1972.

McBurney, D. H., & Bartoshuk, L. M. Interactions between stimuli with different taste qualities. *Physiology & Behavior,* 1973, *10,* 1101–1106.

McBurney, D. H., & Shick, T. R. Taste and water taste of twenty-six compounds for man. *Perception & Psychophysics,* 1971, *10,* 249–252.

McBurney, D. H., Smith, D. V., & Shick, T. R. Gustatory cross-adaptation: Sourness and bitterness. *Perception & Psychophysics,* 1972, *11,* 228–232.

Meiselman, H. L. Human taste perception. *CRC Critical Reviews in Food Technology,* April 1972, pp. 89–119.

Melzack, R., & Wall, P. D. On the nature of cutaneous sensory mechanisms. *Brain,* 1962, *85,* 331–356.

Mozell, M. M., Smith, B. P., Smith, P. E., Sullivan, R. L., Jr., & Swender, P. A technique to occlude the nasal chemoreceptors during lingual flavor stimulation. *Physiology & Behavior,* 1969, *4,* 131.

Murchison, C. (Ed.). *A history of psychology in autobiography* (Vol. 1). Worcester, Mass.: Clark University Press, 1930.

Myers, C. S. The taste-names of primitive peoples. *British Journal of Psychology,* 1904, *1,* 117–126.

Nafe, J. P. The psychology of felt experience. *American Journal of Psychology,* 1927, *39,* 367–389.

Nowlis, G. H., & Frank, M. Qualities in hamster taste: Behavioral and neural evidence. In J. LeMagnen & P. Macleod (Eds.), *Olfaction and taste VI.* London: Information Retrieval, 1977.

Öhrwall, H. Untersuchungen uber den geschmackssinn. *Skandinavisches Archiv für Physiologie,* 1891, *2,* 1–69.

Öhrwall, H. Die modalitäts—Und qualitätsbegriffe in der sinnesphysiologie und deren bedeutung. *Skandinavisches Archiv für Physiologie,* 1901, *11,* 245–272.

O'Mahony, M., & Muhiudeen, H. A preliminary study of alternative taste languages using qualitative description of sodium chloride solutions: Malay versus English. *British Journal of Psychology,* 1977, *68,* 275–278.

Pfaffmann, C. Gustatory afferent impulses. *Journal of Cellular and Comparative Physiology,* 1941, *17,* 243–258.

Pfaffmann, C. The afferent code for sensory quality. *American Psychologist,* 1959, *14,* 226–232. (a)

Pfaffmann, C. The sense of taste. In J. Field, H. W. Magoun, & V. E. Hall (Eds.), *Handbook of physiology: Neurophysiology* (Vol. 1). Baltimore, Md.: Williams & Wilkins, 1959. (b)

Piéron, H. Recherches sur les lois de variation des temps de latence sensorielle en fonction des intensités excitatrices. *L'Année Psychologique,* 1914, *20,* 17–96.

Schiffman, S. S. Physiochemical correlates of olfactory quality. *Science,* 1974, *185,* 112–117.

Schiffman, S. S., & Erickson, R. P. A theoretical review: A psychophysical model for gustatory quality. *Physiology & Behavior,* 1971, *7,* 617–633.

Sekuler, R. Spatial vision. *Annual Review of Psychology,* 1974, *25,* 195–232.

Shore, L. E. A contribution to our knowledge of taste sensations. *Journal of Physiology,* 1892, *13,* 191–217.

Skramlik, E. von. *Handbuch der physiologie der niederen sinne.* Leipzig, Germany: Georg Threme, 1926.

Smith, D. V. Taste intensity as a function of area and concentration. *Journal of Experimental Psychology,* 1971, *87,* 163–171.

Smith, D. V., & Frank, M. Cross adaptation between salts in the chorda tympani nerve of the rat. *Physiology & Behavior,* 1972, *8,* 213.

Smith, D. V., & McBurney, D. H. Gustatory cross-adaptation: Does a single mechanism code the salty taste? *Journal of Experimental Psychology,* 1969, *80,* 101–105.

Stevens, S. S. Sensory scales of taste intensity. *Perception & Psychophysics,* 1969, *6,* 302–308.

Stevens, S. S. *Psychophysics: Introduction to its perceptual, neural, and social prospects.* New York: Wiley, 1975.

Uttal, W. R. *The psychobiology of sensory coding.* New York: Harper & Row, 1973.

Well, A. D. The influence of irrelevant information on speeded classification tasks. *Perception & Psychophysics,* 1971, *10,* 79–84.

Wicker, F. W. Mapping the intersensory regions of perceptual space. *American Journal of Psychology,* 1968, *81,* 178–188.

Woodworth, R. S. The puzzle of color vocabularies. *Psychological Bulletin,* 1910, *7,* 325–334.

Zotterman, Y. Thermal sensations. In J. Field, H. W. Magoun, & V. E. Hall (Eds.), *Handbook of physiology: Neurophysiology* (Vol. 1). Baltimore, Md.: Williams & Wilkins, 1959.

Received October 17, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 168–174

Tests of Significance in Stepwise Regression

Leland Wilkinson
University of Illinois at Chicago Circle

Tests of significance of the sample squared multiple correlation (R^2) in stepwise multiple regression have not been possible because its distribution is unknown. The present study used Monte Carlo simulation and least squares smoothing to construct tables of the upper 95th and 99th percentage points of the sample R^2 distribution in forward selection. A survey of published psychological research that used stepwise regression found a substantial inflation of reported significance levels when compared to the tabled values. Recommendations are given for use of these tables in evaluating results from forward selection and other stepwise methods.

Stepwise regression has a controversial role in statistical data analysis. Since the introduction of various automated techniques for selecting the "best" subset of a set of predictor variables in a multiple regression researchers have been warned about their indiscriminate use (e.g., Kupper, Stewart, & Williams, 1976; Brandt, Note 1). The primary reason for this caution is that for any subset selection procedure based on inspection of the sample data, the usual F statistic for testing the significance of the multiple correlation is biased (Pope & Webster, 1972). Unfortunately, the most widely used computer programs print this statistic at each step without any warning that it does not have the F distribution under automated stepwise selection (Armor & Couch, 1972; Barr, Goodnight, Sall, & Helwig, 1976; Dixon, 1975; Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). Researchers encouraged by a significant multiple correlation from a stepwise analysis are often surprised to find how much it shrinks under cross-validation (Schmitt, Coyle, & Rauschenberger, 1977). For one solution to this problem, the present study used Monte Carlo methods to provide tables of percentage points of the distribution of the sample stepwise squared multiple correlation (R^2) under the null hypothesis that the population multiple correlation is zero. The tabled values can be substituted directly for those on the computer printout when forward selection is used, and they can serve as approximations when other subset selection methods are used.

Distribution of the Sample R² *in Stepwise Regression*

Given a sample of size n from $k + 1$ multivariate normal variates, the sample R^2 between the first and the remaining k variates under the null hypothesis has the beta distribution. In this case, the transformation

$$F = R^2(n - k - 1)/(1 - R^2)k \quad (1)$$

has the F distribution with degrees of freedom k and $n - k - 1$. These distributions apply whether the values of the k predictors are fixed or vary across samples, as long as the null hypothesis is true.

Complications arise when k predictors are chosen from m on the basis of the sample

This research was supported by funds from the Computer Center, University of Illinois at Chicago Circle. A listing of the FORTRAN subroutines is available from the author.

Nancy Hirschberg prompted this research in a discussion of a stepwise regression analysis. Rowell Huesmann, Mike Levine, Jeremy Shefner, and Ron Golland provided valuable advice on the analysis. Bruce Korth, Jerry Dallal, and Shari Diamond commented helpfully on the present article.

Requests for reprints should be sent to Leland Wilkinson, Department of Psychology, University of Illinois, Box 4348, Chicago, Illinois 60680.

data. Two straightforward cases occur, however: when $k = m$ and when $k = 1$. If $k = m$, the above distributions obviously apply. If $k = 1$, and if the predictor that maximizes the sample R^2 is chosen, then the sample R^2 has an independent beta extreme-value distribution. For this case, the F statistic in Equation 1 may be used with critical value

$$\alpha^* = 1 - (1 - \alpha)^{1/m}, \qquad (2)$$

where α is the family critical level.

Beyond these two cases, no exact distributions are known. When the subset of size k $(1 < k < m)$ is chosen to maximize the sample R^2, the distribution of the sample R^2 is an extreme-value distribution of a set of dependent beta variates. This approaches the independent extreme-value distribution when m is large and k is small asymptotically. When the predictors are correlated, this asymptotic convergence is slower, since dependencies among the sample R^2 values for all $\binom{m}{k}$ subsets are greater. Most stepwise algorithms do not necessarily maximize the sample R^2, however, and the distributions in these cases would be even more complicated whether or not the predictors were mutually independent.

Diehr and Hoflin (1974) simulated the sample R^2 distribution for the best R^2 subset among $\binom{m}{k}$ given independence among the predictors. This distribution is particularly important because its percentage points provide an upper bound on sample R^2 values from any subset selection method on independent or correlated predictors. For each $m = 2$–8, $k = 1$–m, and $n = 10, 25, 50, 100$, and 200, they computed 100 R^2 values. Each value was obtained by computing all possible regressions among $\binom{m}{k}$ and selecting the one with the largest R^2. The computing time for this task limited the number of replications and parameter values, but they were able to provide a function that approximates their Monte Carlo results:

$$R^2(k, m, n, \alpha) = w(1 - v^k), \qquad (3)$$

with w and v determined iteratively from the known values of $R^2(1, m, n, \alpha)$ and $R^2(m, m, n, \alpha)$. Furnival and Wilson (1974) developed a rapid "leaps and bounds" algorithm for computing the best subset R^2 that can be used for extending Diehr and Hoflin's results.

Rencher and Fu-Ceayong (Note 2) used Monte Carlo simulation to compute upper percentage points of the sample R^2 distribution in stepwise selection and elimination (Draper & Smith, 1966, p. 171). They generated both uncorrelated and correlated predictors for selection. As in the Diehr and Hoflin (1974) study, their results were approximations from a relatively small number of replications (200–400), but they included a wider range of parameter values.

The present study used a simple algorithm for choosing subsets: forward selection (Draper & Smith, 1966, p. 169). This method is faster than other selection procedures and exceeded many of them in a Monte Carlo study that involved several cross-validation criteria (Dempster, Schatzoff, & Wermuth, 1977). Furthermore, forward selection is the standard or most basic option in most widely used statistical programs for stepwise regression.

Method

Because of the computing time needed for estimating each point in the distributions by Monte Carlo simulation (between 10 sec and 5 min.), a four-stage procedure was used: (a) initial approximation of values in order to select suitable nodes for Monte Carlo estimation, (b) Monte Carlo simulation of the selected nodes, (c) smoothing of the Monte Carlo estimates by graphical and least squares methods, and (d) testing fitted values by new Monte Carlo simulation. Since most of the tabled values were not initially estimated by Monte Carlo, a check on the smoothing process itself was possible in the last stage.

Approximation

The iterative function (Equation 3) was used to generate values of $R^2(k, m, n, \alpha)$ for $k = 1$–m; $m = 2$–30; $n - m - 1 = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150$, and 200; and $\alpha = .05$ and .01. From these values, suitable nodes were selected for fitting nomographs in the two tables to be constructed. These nodes were the known values at $k = 1$ and $k = m$ for $n - m - 1 = 10, 20, 50, 100$, and 200, plus the subsets $(k, m) = (2, 3), (3, 4), (2, 7), (4, 7), (3, 12), (7, 12), (3, 20), (7, 20)$, and $(12, 20)$ at the same degrees of freedom.

Table 1
Upper 95th Percentage Points of Distribution of Sample Squared Multiple Correlation in Forward Selection

		$n - m - 1$															
m	k	10	12	14	16	18	20	25	30	35	40	50	60	80	100	150	200
2	1	38	33	29	26	24	22	18	15	13	12	9	8	6	5	3	2
2	2	45	39	35	31	28	26	21	18	16	14	11	10	7	6	4	3
4	1	39	35	31	28	26	24	20	17	15	14	11	9	7	6	4	3
4	2	50	44	40	37	34	31	26	23	20	18	14	12	9	7	5	4
4	3	55	49	45	41	37	34	29	25	22	19	16	13	10	8	5	4
4	4	58	52	47	43	39	36	31	26	23	21	17	14	11	9	6	5
6	1	38	34	31	29	26	25	21	18	16	14	12	10	8	6	4	3
6	2	50	45	41	38	35	33	28	25	22	20	16	14	11	9	6	5
6	3	57	51	47	43	40	38	32	28	25	22	19	16	12	10	7	6
6	4	61	55	51	47	43	40	34	30	27	24	20	17	13	11	7	6
6	6	66	60	55	51	47	44	37	33	29	26	22	18	14	12	8	6
8	1	36	33	30	28	26	24	21	18	16	15	12	11	8	7	5	4
8	2	49	45	41	38	36	34	29	26	23	21	18	15	12	10	7	6
8	3	57	52	48	45	42	39	34	30	27	24	20	18	14	11	8	7
8	4	62	57	52	49	45	43	37	33	29	26	22	19	15	12	9	7
8	6	68	62	57	53	50	47	40	36	32	29	24	21	16	13	10	7
8	8	71	66	61	56	53	49	43	38	34	30	25	22	17	14	10	7
10	1	35	32	29	27	26	24	21	18	16	15	13	11	8	7	5	4
10	2	48	44	41	38	36	34	30	26	24	21	18	16	12	10	7	6
10	3	56	52	48	45	42	40	35	31	28	25	22	19	15	12	9	7
10	4	61	57	53	50	47	44	38	34	31	28	24	21	16	14	10	8
10	6	69	64	59	55	52	49	43	38	34	31	26	23	18	15	11	9
10	8	72	67	63	59	55	52	45	40	36	33	28	24	19	16	11	9
10	10	75	70	65	61	57	54	47	42	38	34	29	25	20	16	11	9
15	1	31	29	27	25	24	23	20	18	16	15	13	11	9	7	5	4
15	2	44	41	39	36	35	33	29	26	24	22	19	16	13	11	8	7
15	3	53	50	47	44	42	40	35	32	29	27	23	20	16	14	10	8
15	4	59	56	53	50	47	45	40	36	33	30	26	23	18	16	11	9
15	6	68	64	60	57	54	52	46	42	38	35	30	26	21	18	13	11
15	8	74	70	66	62	59	56	50	45	41	38	33	29	23	20	14	11
15	10	78	73	69	66	62	59	53	48	43	40	34	30	24	21	15	11
15	15	81	77	73	69	65	62	56	50	46	42	36	31	25	21	15	11
20	1	27	26	24	23	22	21	19	17	16	14	12	11	9	7	5	4
20	2	40	38	36	34	33	31	28	25	23	21	18	16	13	11	8	6
20	3	50	47	45	42	40	38	35	32	29	27	23	21	17	14	10	8
20	4	57	54	51	48	46	44	40	36	33	31	27	24	19	16	12	10
20	6	66	63	60	57	54	52	47	43	39	37	32	28	23	20	14	12
20	8	73	69	66	63	60	57	52	47	44	40	35	31	26	22	16	13
20	10	78	74	70	67	64	61	56	51	47	43	38	33	27	23	17	14
20	15	85	81	77	74	70	67	61	55	51	47	41	37	30	25	18	14
20	20	85	81	77	74	71	68	62	56	52	48	42	37	30	25	18	14

Note. Decimals are omitted; m = number of predictors; k = number of predictors selected; n = sample size.

Data Generation

Two FORTRAN subroutines were written for the Monte Carlo simulation. The first generated a sample correlation matrix directly from a standardized Wishart distribution given m and n and using an algorithm from Odell and Feiveson (1966). The second subroutine used the abbreviated Gauss-Doolittle method with forward selection to compute a sample R^2 value (Draper & Smith, 1966, p. 178). Sample R^2 values were sorted and noted at the 95th and 99th percentage points after 500 replications. Additional replications in blocks of 100 were continued in each run until a stopping rule was satisfied. The stopping rule was that the upper 99th percentile value be bound on either side by a value differing from it by lesss than .01. Computations were done in double precision on an IBM 370/158 using the FORTRAN H extended compiler and the sort (KB01AD) and normal random number (FA03A) routines from the *Harwell Subroutine Library* (1973).

Table 2
Upper 99th Percentage Points of Distribution of Sample Squared Multiple Correlation in Forward Selection

		$n-m-1$															
m	k	10	12	14	16	18	20	25	30	35	40	50	60	80	100	150	200
2	1	53	47	42	38	35	32	27	23	20	18	14	12	9	8	5	4
2	2	60	54	48	44	40	37	31	26	23	21	17	14	11	9	6	5
4	1	52	47	42	39	36	33	28	24	22	19	16	14	10	9	6	4
4	2	63	57	51	47	44	41	35	30	27	24	20	17	13	11	7	5
4	3	69	62	56	51	48	44	38	33	29	26	22	19	15	12	8	6
4	4	71	64	59	54	50	47	40	35	31	28	23	20	15	12	8	6
6	1	49	45	41	38	36	33	28	25	22	20	17	14	11	9	6	5
6	2	61	56	52	48	45	42	36	32	29	26	22	19	15	12	8	7
6	3	68	62	57	54	50	47	40	35	32	29	24	21	16	14	9	8
6	4	72	66	61	57	53	49	43	38	34	30	26	22	17	14	10	8
6	6	76	71	66	61	57	54	47	41	37	33	28	24	19	15	10	8
8	1	47	43	40	37	35	32	28	25	22	20	17	14	11	9	6	5
8	2	60	55	51	48	45	42	37	33	29	27	23	20	15	13	9	7
8	3	67	62	58	54	51	48	42	37	33	30	26	22	18	15	10	8
8	4	72	66	62	58	54	51	45	40	36	32	27	24	19	16	11	9
8	6	78	72	67	63	59	55	48	43	39	35	30	26	21	17	12	9
8	8	80	75	70	66	62	59	52	46	41	37	32	27	22	18	12	9
10	1	44	41	38	36	34	32	28	24	22	20	17	15	12	9	7	5
10	2	58	54	50	47	44	42	37	33	30	27	23	20	16	13	9	7
10	3	66	61	57	54	51	48	42	38	34	31	27	23	18	15	11	9
10	4	71	66	62	58	55	52	46	41	37	34	29	25	20	17	12	10
10	6	77	72	67	63	60	57	50	45	40	37	32	28	22	19	14	11
10	8	81	76	71	67	63	60	53	47	43	39	33	29	23	20	14	11
10	10	83	78	74	70	66	63	56	50	45	41	35	30	24	20	14	11
15	1	39	36	34	32	31	29	26	23	21	19	17	15	12	10	7	5
15	2	53	50	47	45	42	40	36	32	29	27	23	20	16	13	9	7
15	3	62	58	55	52	50	47	42	38	35	32	27	24	19	16	11	9
15	4	68	64	60	57	55	52	47	42	38	35	30	27	22	18	13	11
15	6	75	71	67	64	61	58	52	47	43	40	34	30	25	21	15	13
15	8	80	76	72	68	65	62	56	51	46	43	37	33	27	23	17	14
15	10	84	80	76	72	69	66	59	53	49	45	39	35	28	24	18	14
15	15	87	83	80	76	73	70	63	57	53	49	42	37	30	25	18	14
20	1	35	33	31	30	28	27	24	22	20	19	16	14	12	10	7	5
20	2	49	47	44	42	40	38	34	31	28	26	22	20	16	13	9	7
20	3	58	55	52	50	48	46	41	37	34	32	27	24	19	16	11	9
20	4	64	61	58	55	53	51	46	42	38	36	31	27	22	19	13	11
20	6	72	69	66	63	60	58	52	48	44	41	36	32	26	22	16	14
20	8	78	74	71	68	65	63	57	52	48	45	39	35	29	24	18	16
20	10	82	79	75	72	69	66	60	55	51	48	42	37	31	26	20	16
20	15	89	86	82	79	76	73	67	61	57	53	46	41	34	29	21	16
20	20	90	87	83	80	77	75	68	63	58	54	48	42	35	29	21	16

Note. Decimals are omitted; m = number of predictors; k = number of predictors selected; n = sample size.

Curve Fitting

The nodes from the Monte Carlo simulation and the known values at $k = 1$ and $k = m$ were used to draw nomographs for $n - m - 1 = 10$–200, $m =$ 2–20, and $k = 1$–m. Initial estimates of the entries in Tables 1 and 2 were read off these nomographs. Most of these values were not estimated directly by Monte Carlo.

To smooth further the results from the nomographs, the 656 entries in each table (including the known values) were predicted by multiple regression. Twenty linear and nonlinear terms plus their interactions were constructed from approximations contained in Diehr and Hoflin (1974), Zirphile (1975), and Kendall and Stuart (1969, p. 330). The standard error of the estimate in predicting the 95th percentile table values from the best 15 terms in

the equation was .0054; for the 99th percentile table, it was .0047. The entries in Tables 1 and 2 were taken from the regression estimates.

Check on Accuracy

As a check on the accuracy of the table values, simulations were run to predict 20 points that had not previously been estimated by Monte Carlo. Runs were continued in blocks of 1,000 replications until a stopping rule was met. The stopping rule was that two cumulative runs result in a successive difference of less than .01 in both the 95th and 99th percentile values. Of the 40 values tested in the combined tables, 17 differed from the Monte Carlo value by .01, and 23 agreed exactly with the printed accuracy. No significant differences in errors were found across tables.

Results

Histograms of the simulated data at selected parameter values resembled beta distributions, although no beta-type function could be found to reproduce accurately the upper tail values. Tables 1 and 2 give the 95th and 99th percentage points of these distributions.

Use of Tables

The tables have been constructed to cover a full range of practical parameter values. Linear interpolation works quite well for m and n. For interpolations on k, however, graphical plotting of the table values provides more accurate estimates. Extrapolation may be used moderately, since the values near the margins of the table change slowly. The known values when m and n are large, for example, can be extrapolated accurately on fine graph paper with a flexible drafting curve.

Discussion

The tables clearly illustrate the inflation of the sample R^2 in stepwise regression that has frequently been noted by statisticians. For example, $R^2(4, 4, 35, .05) = .26$, whereas $R^2(4, 20, 35, .05) = .51$. This inflation has not always been noted by researchers, however. A computer-assisted search for articles in psychology using stepwise regression from 1969 to 1977 located 71 articles. Out of these articles, 66 forward selection analyses reported as significant by the usual F tests were found. Of these 66 analyses, 19 were not significant by Table 1.

The extent of this artifact may have contributed to the poor reputation of subset selection methods in multiple regression through failures to replicate published research. This situation is ironic because of the clear evidence of the substantial superiority of forward selection over ordinary least squares in a variety of prior distributions. Forward selection can be almost as effective as ridge regression in minimizing prediction and beta weight errors with highly correlated predictors (Dempster et. al., 1977). Furthermore, forward selection offers a more parsimonious model than does ridge regression because only k predictors out of m are included in the equation.

Three questions remain, however, regarding the application of these tables: (a) Can the tables be used for other subset selection methods? (b) Can they be used when predictors are intercorrelated? (c) Do they apply when k is not known prior to the analysis?

In answer to the first question, the tabled values may be compared to results from Monte Carlo studies of two other subset selection methods: best subset (Diehr & Hoflin, 1974) and stepwise selection and elimination (Rencher & Fu-Ceayong, Note 2). Diehr and Hoflin's approximation, given in Equation 3, yields values higher than those in Tables 1 and 2. The differences range from .01 when $m = 3$ to .10 when $m = 20$. This discrepancy is due partly to the approximation and partly to the fact that the sample R^2 percentage point for the best subset case is an upper bound for all subset R^2 percentage points at the same parameter values. Rencher and Fu-Ceayong's results fit the values within both tables closely. For 15 comparable parameter values in both tables, the largest discrepancy was .02, with most values less than .01. This fit indicates that the tables should be appropriate for stepwise selection and elimination. Although the various other suboptimal selection methods used

in most computer programs occasionally result in different subsets of size k from $\binom{m}{k}$, the distribution of the sample R^2 in these cases may nevertheless fit these tables, particularly when m is large and k is small.

The question of intercorrelated predictors requires further Monte Carlo simulation for an answer. Rencher and Fu-Ceayong investigated this problem using stepwise regression on random predictors intercorrelated in various ways. The upper percentage points of the sample R^2 distributions from correlated predictors were only slightly lower than those for independent predictors in all cases. The results for forward selection should be similar. In any case, loss of power should not be substantial when these tables are used for correlated predictors, provided k is small relative to m.

Finally, when k is unknown prior to the analysis, stopping rules must be applied (Bendel & Afifi, 1977). In this case, k is a random variable instead of a fixed constant. Further Monte Carlo research is needed to identify the effect of stopping rules on the sample R^2 distribution. The most common stopping rule for forward selection is to continue stepping until the sample partial correlation is "nonsignificant" by a standard F test (Draper & Smith, 1966, p. 71). As an alternative, simultaneous inference may be used for subset selection to control the Type I family error rate (Aitkin, 1974). This method is conservative, however, like most simultaneous test procedures; for a given critical level, it will eliminate fewer subsets containing "unimportant" predictors than will forward selection with a sequential stopping rule.

Problems remain regarding tests of significance of coefficients in stepwise regression. In lieu of such tests, cross-validation should be performed, even though the standard errors of the coefficients may be smaller than those in the corresponding ordinary least squares equation on m variables. Researchers should carefully consider the advantages and disadvantages of various subset selection methods and other biased estimation methods for analyzing particular data sets (for reviews, see Hocking, 1976, and Jennrich, 1977). Users of standard, automated stepwise computer programs, however, should choose forward selection, ignore the tests of significance printed at each step, and consult Tables 1 and 2 to evaluate the significance of the final equation they select.

Reference Notes

1. Brandt, E. B. *The abuse of stepwise regression or: Experiments in fishing* (Rand Corp. P-4260). Santa Monica, Calif.: Rand Corporation, 1970.
2. Rencher, A. C., & Fu-Ceayong, P. *Inflation of* R^2 *in best subset regression.* Unpublished manuscript, Brigham Young University, Department of Statistics, 1977.

References

Aitkin, M. A. Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics,* 1974, *16,* 221–227.

Armor, D. J., & Couch, A. S. *Data-Text primer: An introduction to computerized social data analysis.* New York: Free Press, 1972.

Barr, A. J., Goodnight, J. H., Sall, J. P., & Helwig, J. T. *A user's guide to SAS.* Raleigh, N.C.: SAS Institute, 1976.

Bendel, R. B., & Afifi, A. A. Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association,* 1977, *72,* 46–53.

Dempster, A. D., Schatzoff, N., & Wermuth, N. A simulation study of alternatives to least squares. *Journal of the American Statistical Association,* 1977, *72,* 77–91.

Diehr, G., & Hoflin, D. R. Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics,* 1974, *16,* 317–320.

Dixon, W. J. (Ed.). *BMDP: Biomedical computer programs.* Berkeley: University of California Press, 1975.

Draper, N. R., & Smith, H. *Applied regression analysis.* New York: Wiley, 1966.

Furnival, G. M., & Wilson, R. W. Regressions by leaps and bounds. *Technometrics,* 1974, *16,* 499–511.

Harwell subroutine library. Harwell, Berkshire, England: U.K. Atomic Energy Authorized Research Group, Theoretical Physics Division, 1973.

Hocking, R. R. Analysis and selection of variables in linear regression. *Biometrics,* 1976, *32,* 1–49.

Jennrich, R. I. Stepwise regression. In K. Enslein, A. Ralston, & H. Wilf (Eds.). *Statistical methods for digital computers* (Vol. 3). New York: Wiley, 1977.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 1). New York: Hafner, 1969.

Kupper, L. L., Stewart, J. R., & Williams, K. A. A

note on controlling significance levels in stepwise regression. *American Journal of Epidemiology,* 1976, *103,* 13–15.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. *SPSS: Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill, 1975.

Odell, P. L., & Feiveson, A. H. A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association,* 1966, *61,* 199–203.

Pope, P. T., & Webster, J. T. The use of an *F*-statistic in stepwise regression procedures. *Technometrics,* 1972, *14,* 326–340.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. *Psychological Bulletin,* 1977, *84,* 751–753.

Zirphile, J. Letter to the editor. *Technometrics,* 1975, *17,* 145.

Received October 20, 1977 ∎

Psychological Bulletin
1979, Vol. 86, No. 1, 175–190

Suppression of Responding During Signaled and Unsignaled Shock

Norman Hymowitz
New Jersey Medical School
College of Medicine and Dentistry of New Jersey

The empirical basis for Seligman's safety signal hypothesis derives largely from studies of response suppression during signaled and unsignaled shock and from studies of animals' preference for signaled over unsignaled shock. Recently, the literature on preference for signaled over unsignaled shock has been the subject of serious criticism and controversy, thus weakening the empirical foundations of the safety signal hypothesis. The present article reviews the literature on response suppression to determine if this, too, has been the subject of controversy and criticism. To the contrary, the suppression literature provides strong support for the safety signal hypothesis and also reports data that are compatible with much of the choice literature. This agreement between the two tests of the safety signal hypothesis increases confidence in the reliability of the data and the adequacy of the hypothesis. Despite this agreement, emerging data on response suppression during signaled and unsignaled shock suggest that, at best, the safety signal hypothesis emphasizes only one of the many determinants of differential response suppression.

Seligman (1968) reported that food-maintained responding was more readily suppressed by a given intensity of electric shock when shock delivery was unsignaled than when it was signaled. Furthermore, animals exposed to unsignaled electric-shock delivery developed significantly more gastric ulcers than did animals exposed to comparable signaled shock. Seligman interpreted these findings in terms of a safety signal hypothesis. According to this hypothesis, the presence and absence of the preshock signal or conditioned stimulus (CS) specify shock ("danger") and shock-free ("safety") occasions, respectively. During signaled shock, the animals spend most of the session in safety. With unsignaled shock, most of the session is spent in danger. Presumably, this differential exposure to danger and safety is responsible for differential response suppression and ulceration during signaled and unsignaled shock.

Since Seligman's report the analysis of the behavioral effects of signaled and unsignaled electric-shock delivery has proceeded in two directions. On the one hand, a host of researchers have attempted to systematically replicate Seligman's (1968) original findings with response suppression as the data of interest. As noted by Seligman and Binik (1977), such studies have uniformly confirmed Seligman's finding of more response suppression during unsignaled than during signaled shock. A second avenue of investigation involved choice of signaled over unsignaled shock delivery as the dependent measure. According to the safety signal hypothesis, animals ought to select signaled over unsignaled electric-shock delivery. Indeed, a host of investigators have suggested that this is the case (e.g., Badia & Culbertson, 1972; Badia, Culbertson, & Harsh, 1973; Harsh & Badia, 1975; Lockard, 1963; Perkins, Seymann, Levis, & Spencer, 1966). However, some negative findings (e.g., Biederman & Furedy, 1973, 1976b; Crabtree & Kruger,

Requests for reprints should be sent to Norman Hymowitz, Behavioral Sciences Unit, Department of Psychiatry and Mental Health Sciences, New Jersey Medical School, 100 Bergen Street, Newark, New Jersey 07103.

1975) and serious methodological criticisms of the choice literature (e.g., Biederman & Furedy, 1976a; Furedy, 1975) exist.

As noted by Biederman and Furedy (1976a), many of the earlier studies of animals' preference for signaled over unsignaled shock used unscrambled grid shock (e.g., Lockard, 1963). Since animals may modify or avoid unscrambled shock by postural adjustment, unambiguous evaluation of such studies is not possible. Moreover, Biederman and Furedy (1973) found preference for signaled over unsignaled shock when the shock was unscrambled but not when it was scrambled.

Two studies reported preference for signaled shock over unsignaled shock when shock was delivered to the animals through electrodes attached directly to the animal's tail (Miller, Daniel, & Berk, 1974; Perkins et al., 1966). Since it is not possible to modify shock delivered through tail electrodes, the data seem to support the safety signal hypothesis. However, Biederman and Furedy (1976a, 1976b) pointed out that the tail electrode procedure was unreliable and that many of the animals were eliminated from the studies because of damage to the electrodes. In view of the unreliable procedure and possible sampling bias, the data may hardly be viewed as strong support for the safety signal hypothesis.

A more recent study by Miller, Marlin, and Berk (1977) that employed tail shock is less susceptible to serious criticism. Miller and his students perfected their apparatus and technique so that they were able to successfully and reliably deliver shock directly to the tail of freely moving animals (Berk, Marlin, & Miller, 1977). Five of the eight animals studied showed consistent preferences for the side of the shuttle box associated with signaled shock. The three other animals revealed an initial preference for signaled shock, but failed to maintain the preference during the course of several reversal conditions in which the side of the shuttle box associated with signaled shock was changed.

Perhaps the most thorough analyses of animals' preference for signaled over unsignaled shock are those that were conducted by Badia and his students (Badia & Culbertson, 1972; Badia et al., 1973; Harsh & Badia, 1975). Badia used a changeover procedure to study the choice behavior of rats. Typically, the animals were exposed to unsignaled electric-shock delivery. Responses on the changeover lever produced, for a brief period of time, a correlated stimulus in the presence of which shocks were preceded by a signal. He also used scrambled grid shock. Hence, his studies are not subject to the criticisms mentioned earlier. However, Biederman and Furedy (1976a, 1976b) criticized Badia's studies on several other counts. According to Biederman and Furedy, Badia typically confounded the correlated stimulus and the CS; that is, the same response that produced the correlated stimulus (light) also produced the preshock signal or CS (tone). Biederman and Furedy (1973) showed that during shock animals pressed to produce a light (correlated stimulus) whether or not shock was preceded by the CS. They suggested that changeover responding was maintained by stimulus change or photic reinforcement, not by signaled shock per se. Biederman and Furedy (1976a, 1976b) also noted that Badia's studies were not balanced. Animals changed over from dark, unsignaled shock to light, signaled shock. If given the opportunity, they might have changed over from light, signaled shock to dark, unsignaled shock.

It is not the purpose of the present article to evaluate the pros and cons of the various discrepancies and controversies in the choice literature (see Badia & Harsh, 1977a, 1977b). Polemics are no substitute for carefully planned experiments. However, closer attention to the first avenue of investigation mentioned above, the suppression of responding during signaled and unsignaled shock, may advance the analysis of the behavioral effects of signaled and unsignaled shock and also may bear upon some of the discrepancies in the choice literature. When, for example, a variable is shown to influence choice of signaled over unsignaled shock and differential response suppression during signaled and unsignaled shock in the same manner, more con-

fidence can be placed in the generality and reliability of the data and in the hypothesis under test (Ghiselin, 1969). Thus, one purpose of the present review of the literature on response suppression during signaled and unsignaled shock is to determine, wherever possible, whether variables purported to influence the choice of signaled shock also influence differential response suppression. Such an analysis should enhance our understanding of the behavioral effects of signaled and unsignaled shock and should suggest additional studies that bear critically on existing controversies in the choice literature.

A second purpose of the present review is to evaluate further the utility of the safety signal hypothesis. It is important to note that the safety signal hypothesis originally was formulated on the basis of findings on differential response suppression (Seligman, 1968). Can the hypothesis accommodate the data that have been generated since 1968? How general are Seligman's (1968) original findings? The following review of the literature provides answers to these questions.

Naturally, all reviews of the literature must be selective. I did not include in the review the growing literature on the somatic effects of signaled and unsignaled shock. For the most part, the focus of research in this area has been upon producing somatic reactions with little attention given to the animal's behavior. Where appropriate, studies from related areas of aversive conditioning are cited and discussed. In particular, studies of choice of signaled over unsignaled shock are presented in detail. To facilitate the comparison of variables that influence choice and suppression in the same manner, the ensuing literature review is organized according to the variables of which differential response suppression is a function.

Experimental Design

Two basic experimental designs have been employed to study differential suppression. With between-groups designs (e.g., Seligman, 1968), the data of interest are comparisons between the rate of responding in the absence of the CS for one group of animals and the rate of responding during comparable unsignaled shock delivery for another group. For within-subject designs (e.g., MacDonald, 1973) the rate of responding in the same animal is studied under signaled as well as unsignaled shock conditions (cf. Sidman, 1960).

MacDonald (1973; MacDonald & Baron, 1973) studied rates of responding in the rat under a two-component multiple schedule in which each component consisted of a two-link chained schedule. Under this schedule, responding in one of the initial links of the chain schedules produced one of the two terminal links during which food reinforcers and either signaled or unsignaled shocks were presented. When signaled and unsignaled electric-shock delivery was scheduled in the separate terminal links, the rats responded at lower rates in the initial and terminal links associated with unsignaled shock. Hymowitz (1976b, 1977b) studied responding within the same animal during multiple and mixed schedules of signaled and unsignaled electric-shock delivery. Much more response suppression occurred in the components of the multiple schedule that were associated with unsignaled shock than in the components associated with comparable signaled shock. Differential response suppression during signaled and unsignaled shock delivery was not obtained when the animals were exposed to the mixed schedule of shock delivery. The significance of the latter finding is discussed in another section.

Although the between-groups and within-subject designs yield compatible results, the within-subject design seems more suitable for parametric analyses. Each animal may differ with respect to the conditions necessary to reveal differential suppression (cf. Hymowitz, 1977b). Factorial designs provide information about parameter values, but do not overcome the differences in sensitivity of individual animals to experimental conditions. A case in point is the present controversy in the literature on the preference for signaled over unsignaled shock.

Several researchers (Biederman & Furedy, 1973, 1976a, 1976b; Crabtree & Kruger, 1975; Furedy & Biederman, 1976) failed to find preferences in animals for signaled over

unsignaled electric-shock delivery. Other investigators (e.g., Badia & Culbertson, 1972; Harsh & Badia, 1976) routinely report positive results. One difference among these researchers is their selection of experimental designs. The former have typically employed between-subjects designs, whereas the latter have studied the behavior of individual animals under a wide range of parameter values. At some parameter values, animals selected signaled shock; at others, they did not. Often, parameters such as shock intensity required adjustment and manipulation during the course of the study to successfully demonstrate preference. Although each of the researchers presented data that bear on our understanding of the preference-of-signaled-shock phenomenon, one wonders whether the difference in the nature of their findings is not in part because of their selection of experimental designs. Within-subject designs are, by definition, better suited for the analysis of phenomena that require constant adjustment of parameter values for individual animals.

Recent studies (Hymowitz, 1977a, 1977b) showed that parametric considerations are of paramount importance in the analysis of differential suppression during signaled and unsignaled shock. These studies are discussed in more detail in later sections, but briefly their results showed (a) that animals differ with respect to shock intensities that yield differential suppression and (b) that for the same animal, a given shock intensity may yield differential suppression of one response but not of another. The failure to adequately take into account individual differences in sensitivity to important controlling parameters may lead to seriously misleading conclusions.

Duration of the CS

Does the duration of the CS in any way affect the differential suppression of responding during signaled and unsignaled shock? No study systematically attempted to answer this question. It is known that CS durations of 60 sec (e.g., Seligman, 1968), 10 sec (Hymowitz, 1977a), and 5 sec (e.g., Hymowitz, 1976b) readily yield differential suppression during signaled and unsignaled shock. In preference situations, it does not seem to matter whether CSs of 3, 5, 10, 20, or 30 sec are employed. Animals readily select the signaled over the unsignaled shock condition (French, Palestino, & Leeb, 1972).

Very likely, the effects of the CS duration on differential response suppression and on preference for signaled over unsignaled shock depend on the relative duration of the CS compared to non-CS occasions. Such a finding would not be inconsistent with Seligman's (1968) safety signal hypothesis. In studies of conditioned suppression (Stein, Sidman, & Brady, 1958), the suppression of responding during the CS depended on the relative durations of CS and non-CS occasions. In brief, when the CS duration was short, relative to between-stimuli durations, responding tended to be more suppressed during the CS than on occasions when the duration of the CS was relatively long. Indeed, Harsh and Badia (1976) employed a CS duration of 30 sec and showed that rats failed to prefer signaled over unsignaled shock when the shock was delivered on the average of once every 45 sec. Preference was observed within the same animals when shock was delivered less frequently. Perhaps the 45-sec shock delivery schedule would have yielded preference for signaled shock if a shorter CS had been employed.

Perkins et al. (1966), on the other hand, failed to find preference for signaled over unsignaled shock when the CS was only .5 sec, although shock was delivered on the average of once every 5 min. They did report preferences for signaled shock with CS durations of 3 and 18 sec. These findings are compatible with Perkins' (1968) preparedness hypothesis. According to this hypothesis, one important reason why animals select signaled over unsignaled shock (or suppress responding more during unsignaled than during signaled shock) is that the CS enables animals to prepare for the impending shock. Presumably, .5 sec is not enough time for such preparation. It would be highly informative to determine if differential response suppression during signaled and unsignaled shock occurs with a brief .5-sec duration CS. An

answer to this question would bear critically on the preparedness hypothesis.

Shock Parameters: Intensity, Duration, and Frequency

Shock Intensity

There is little doubt that the suppression of food-maintained responding generally is directly related to shock intensity. The higher the intensity, the greater is the suppression (cf. Azrin & Holz, 1966; Church, 1969; D'Amato, 1970). Similarly, the suppression of responding in the presence of the CS in the conditioned suppression paradigm also is a function of shock intensity (cf. Kamin, 1965).

Seligman and Meyer (1970) studied the effects of shock intensity on responding during signaled and unsignaled shock. Using a between-groups design and two shock intensities, .60 and 1.00 mA, they showed (a) that for both shock intensities, more response suppression occurred in the groups of rats exposed to unsignaled shock; (b) that responding during unsignaled, mild (.60 mA) shock recovered considerably over sessions; and (c) that little if any recovery during unsignaled 1.0 mA shock occurred.

Hymowitz (1977b) extended these findings. He conducted a within-subject analysis of shock intensity in which each animal was systematically subjected to a wide range of shock intensities. Lever pressing was studied first during 20–25 sessions of multiple fixed-ratio 10 fixed-ratio 10 (multiple FR 10 FR 10) (Rat 1) and multiple FR 20 FR 20 (Rats 2 and 3) food delivery schedules. A multiple variable-time 120-sec variable-time 120-sec (multiple VT 120-sec VT 120-sec) schedule of shock delivery was then superimposed upon the food schedule. In one component of the multiple schedule, a 5-sec tone preceded each shock (signaled shock). In the other component, the preshock stimulus was not presented (unsignaled shock). Blocks of 5–10 successive food-and-shock sessions at a constant shock intensity alternated with blocks of 3–5 successive food-alone sessions. Each component was 6 min. in duration, and each daily session terminated after eight components.

The major findings of this study may be summarized as follows: (a) Each animal differed with respect to the intensities of shock that led to differential response suppression; (b) relatively mild intensities of shock had little effect on responding in the components of the multiple schedule associated with signaled and unsignaled shock, intermediate intensities disrupted responding primarily in the component associated with unsignaled shock, and more severe intensities disrupted responding in both components; and (c) continued testing under signaled and unsignaled shock delivery affected the minimum intensity of shock required to produce differential suppression. Originally, .30 mA shock failed to yield differential response suppression in two of the animals. Following additional testing under higher intensities of shock, signaled and unsignaled shock intensities of .20 mA and .30 mA readily led to more response suppression during unsignaled than during signaled shock in each of the animals.

It should be noted that Hymowitz's (1977b) suppression data are in fairly close agreement with choice data reported by Harsh and Badia (1976). In Badia's (e.g., Badia & Culbertson, 1972; Harsh & Badia, 1976) studies, the animals were exposed to unsignaled response-independent electric-shock delivery. Responses on a changeover lever produced for a brief period of time a correlated stimulus in the presence of which a brief tone preceded shock delivery. Harsh and Badia (1975) reported that changeover responding from unsignaled to signaled shock was a function of shock intensity. At intensities of shock of less than .60 mA, little changeover responding occurred initially. As the intensity increased, the animals spent more time in the changeover condition. Moreover, when the intensity was reduced to .40 mA, changeover responses still occurred at a high rate.

Duration of Shock

Like shock intensity, the duration of shock is directly related to the suppression of re-

sponding. For a given shock intensity, the longer the duration, the greater is the response suppression (e.g., Church, Raymond, & Beauchamp, 1967). Only MacDonald (1973) and MacDonald and Baron (1973) employed shock durations of longer than .50 sec. They used a duration of 2.00 sec and found more suppression in the links of a multiple chain schedule associated with unsignaled shock than in the links associated with signaled shock. It is unfortunate that more systematic data on the effects of shock duration on differential suppression are not available. Studies of choice of signaled over unsignaled shock (Badia et al., 1973) suggest that changeover from unsignaled to signaled shock is a function of shock duration. At durations of signaled shock beyond 2.00 sec, animals ceased changeover responding from briefer unsignaled shock.

It is interesting to note that studies that failed to find preferences for signaled over unsignaled scrambled shock delivery used shock durations of 5.00 sec (Biederman & Furedy, 1973, 1976b; Furedy & Biederman, 1976) and 2.00 sec (Crabtree & Kruger, 1975). Studies that showed preference for signaled and unsignaled scrambled shock delivery (e.g., Badia & Culbertson, 1972; Hymowitz, 1973c) used brief .50-sec shock durations. Parametric studies of the effects of shock duration on response suppression during signaled and unsignaled shock delivery may better describe the relationship between shock duration and responding during signaled and unsignaled shock.

Shock Frequency

The frequency of shock delivery, like the duration and intensity of shock, generally is directly related to response suppression. The more frequent the shock, the greater is the suppression of responding (cf. Azrin & Holz, 1966). The frequencies of shock delivery under which differential response suppression during signaled and unsignaled shock delivery has been found range from an average of one shock every 11 min. (Davis, Memmott, & Hurwitz, 1976) to one every 4 min. (MacDonald & Baron, 1973), 2 min. (Hymowitz, 1976a), 1 min. (Hymowitz, 1973a), and 45 sec (Imada & Okamura, 1975). Although shock frequency undoubtedly influences differential suppression, it is clear that differential response suppression occurs under a wide range of shock frequencies.

It is of some interest that preference for signaled over unsignaled shock is affected by the frequency of shock delivery. If shock delivery is too frequent (i.e., once every 45 sec), little if any preference for signaled shock is found (Harsh & Badia, 1976). In terms of Seligman's (1968) safety signal hypothesis, the failure to obtain preference at relatively short intershock intervals suggests that some minimum period of shock-free time is required to maintain changeover responses.

Again, an analysis of the effects of relatively dense and lean schedules of shock delivery on differential response suppression would be informative. Such an analysis might accompany an analysis of shock intensity; that is, for each shock frequency, one would determine the range of shock intensities that (a) mildly affects responding during signaled and unsignaled shock, (b) suppresses responding primarily during unsignaled shock, and (c) severely suppresses responding under signaled as well as unsignaled shock. Very likely, the shock intensities required to produce these effects would vary as a function of shock frequency. It is an empirical question, however, whether some minimum period of safety is necessary to produce differential response suppression during signaled and unsignaled shock.

Response-Independent and Response-Dependent Electric-Shock Delivery

Most of the studies of the suppressive effects of signaled and unsignaled shock employed response-independent shock delivery. Two studies (Hymowitz, 1976b; MacDonald, 1973) used response-dependent shock delivery schedules. Although it is not possible to discuss the relative contribution of the contingency between shock and responding to differential response suppression on the basis of the available literature, it is clear that each mode of shock presentation is conducive to differential suppression.

It is surprising that more data on this variable are not available. Although there is some question as to whether responding is suppressed more by response-dependent than by response-independent shock delivery (e.g., Azrin, 1956; Camp, Raymond, & Church, 1967; Church, 1969; Church, Wooten, & Matthews, 1970; Hoffman & Fleshler, 1965; Orme-Johnson & Yarczower, 1974; Rachlin & Herrnstein, 1969), it is generally accepted that responding in the absence of the CS in conditioned suppression studies is more affected by response-independent than by response-dependent shock that is delivered in the presence of the CS (Hoffman & Fleshler, 1965; Hunt & Brady, 1955; Orme-Johnson & Yarczower, 1974). Such differences between the two modes of shock delivery might be expected to influence the differential suppression of responding during signaled and unsignaled shock.

Schedule of Electric-Shock Delivery

Only one study (Hymowitz, 1973a) compared the rate of responding during signaled and unsignaled shock delivery during different schedules of shock presentation. For both fixed-interval and variable-interval shock delivery schedules, more suppression of baseline responding was found for animals exposed to unsignaled than to signaled shock. There was little if any difference in suppression between the signaled fixed- and variable-interval conditions.

When the percentage of occasions on which the CS preceded shock was decreased from 100% to 50%, the importance of the schedule of shock delivery was apparent. Much more suppression of baseline responding was found in the 50% variable-interval than in the 50% fixed-interval condition. For the variable-interval shock schedules, as much or more suppression occurred in the 50% as in the unsignaled shock condition (0%). For the fixed-interval shock delivery schedule, the most suppression occurred during unsignaled shock, the least during 100% signaled shock, and an intermediate amount during 50% signaled shock. These findings are discussed further in the section entitled Predictability of Shock. Considering the importance of the schedule of shock delivery for the suppression of responding in other situations (e.g., Azrin, 1956; Camp, Raymond, & Church, 1966; Ferraro, 1967; Morse & Kelleher, 1966), it is not surprising that the schedule interacts with the predictability of shock to influence the course and degree of response suppression.

Appetitive Factors: Food Schedule and Food Deprivation

Schedule of Food Delivery

The schedule and frequency of food delivery influence the manner in which responding is suppressed by noxious events (e.g., Azrin & Holz, 1966). Relatively little is known about their influence on differential response suppression during signaled and unsignaled shock, although it is expected that the frequency and schedule of food delivery will interact with the shock delivery conditions.

In terms of the generality of the effects of signaled and unsignaled shock delivery, it is noteworthy that Hymowitz found differential suppression when food pellets were delivered under variable-interval (Hymowitz, 1976b), fixed-ratio (Hymowitz, 1977b), and fixed-interval (Hymowitz, 1977a) schedules. Imada and Okamura (1975) studied the effects of signaled and unsignaled shock delivery on operant licking in water-deprived animals. Licking was reinforced on a continuous reinforcement schedule. More suppression of licking was found during unsignaled than during signaled shock conditions. These data certainly enhance the generality of Seligman's (1968) original findings obtained with variable-interval food delivery schedules.

Although direct comparisons among the different food schedules used in the abovementioned studies are not possible, it is noteworthy that the manner in which responding is affected by signaled and unsignaled shock is influenced by the schedule of food delivery. With fixed-ratio food delivery, for example, shock led to increases in the length of the postpellet pause, but had little effect on the running rate until much more intense shocks were delivered (Hymowitz, 1977b). For a

given shock intensity, the pausing was much more apparent in the component of the multiple schedule associated with unsignaled shock. At some intensities, little if any pausing occurred during signaled shock, whereas marked pausing occurred during unsignaled shock.

Food Deprivation

As noted by Millenson and de Villiers (1972), the contribution of food deprivation to response suppression, though very important, is a relatively neglected area of research. This is somewhat surprising because food deprivation plays an important role in motivational analyses of punishment (Estes, 1969) and conditioned suppression (Millenson & de Villiers, 1972). Moreover, there is little doubt that this variable interacts with the schedule and frequency of food delivery to determine the resistance of responding to the suppressive effects of electric-shock delivery. Future analyses of response suppression during signaled and unsignaled shock ought to include this variable in the design of the experiment. At present, virtually no data on the effects of food deprivation on differential suppression during signaled and unsignaled shock have been published, although the suppression of responding in the presence of the CS is influenced by food deprivation (Millenson & de Villiers, 1972).

Dependent Variable

Myers (1971) noted that too much of our knowledge about response suppression derives from studies in which operant lever pressing served as the dependent variable. He questioned the generality of the lever-press findings to other response systems. Studies of autoshaping, learned taste aversions, species-specific defense reactions, and schedule-induced behaviors (cf. Bolles, 1972; Seligman, 1970) produced data that seriously question the generality of principles of learning derived primarily from analyses of key pecking in pigeons and lever pressing in rats. One may similarly question the generality of the data on signaled and unsignaled shock, since most investigators employed schedule-controlled lever pressing as the dependent variable. Some data recently collected with schedule-induced water intake as the dependent variable serve to extend the generality of the lever-press findings (Hymowitz, 1977a).

Schedule-induced water intake was first reported by Falk (1961) and refers to post-pellet water intake that occurs when food-deprived rats, mice, pigeons, or monkeys are exposed to any one of a number of intermittent schedules of food delivery (Falk, 1971; Segal, 1972; Staddon, 1975, 1977a, 1977b). Induced water intake may be distinguished from normal regulatory drinking. With schedule-induced drinking, the animals are not water deprived, and the amount of water consumed may be far in excess of physiological requirements (polydipsia). Induced drinking may also be distinguished from paradigms in which licking serves as an operant on which the delivery of food is dependent. Schedule-induced licking follows pellet delivery and persists even though licking may postpone the delivery of the food pellet (Falk, 1971).

An adequate explanation of schedule-induced water intake must take into account that drinking is but one of a number of responses that animals may engage in following pellet ingestion. Other schedule-induced behaviors are schedule-induced wheel running, pica, escape, air licking, and attack (Falk, 1971). A common feature of each behavior is that they occur during those occasions when the probability of food delivery is low. Thus, they may be related to aversive aspects of intermittent positive-reinforcement schedules. Falk (1971) also noted that the induced behaviors seem not to be related to the task at hand—the procurement of food pellets. He termed these behaviors *adjunctive* behaviors and likened them to displacement activities more commonly studied by ethologists. Staddon (1975, 1977a, 1977b) distinguished between *interim* behaviors, such as schedule-induced water intake and grooming, which occur during occasions when food pellets are not available, and *terminal* behaviors, such as lever pressing, which occur when food pellets are available. Interactions between the two behaviors may provide the

underlying structure for schedule performances observed in the operant chamber.

Although the exact nature of schedule-induced water intake eludes researchers at present, schedule-induced water intake may serve as a useful dependent variable for the analysis of the behavioral effects of signaled and unsignaled electric-shock delivery. Schedule-induced licking shares with schedule-controlled lever pressing many features of a desirable dependent variable. It is a highly quantifiable sample of behavior that remains quite stable from session to session. Moreover, it also is sensitive to variables such as the frequency of food delivery (Falk, 1967; Freed & Hymowitz, 1972), the magnitude of the food pellet (Falk, 1967; Freed & Hymowitz, 1972), body-weight loss (Freed & Hymowitz, 1972), shifts in the rate of reinforcement (Jacquet, 1972), and the delivery of electric shock (e.g., Hymowitz & Freed, 1974). Comparisons of the effects that various conditions of signaled and unsignaled shock have upon the suppression of schedule-controlled and schedule-induced behavior may provide useful insight into the role that the nature of the response has in the suppression of behavior in general.

Recent studies (Bond, Blackman, & Scruton, 1973; Dunham, 1971; Freed, Hymowitz, & Fazzaro, 1974; Hymowitz, 1973b, 1976a; Hymowitz & Freed, 1974) showed that schedule-induced licking was suppressed by shock in very much the same manner as lever pressing. However, the intensity of shock required to suppress induced licking was lower than the intensity required to suppress lever pressing (Bond et al., 1973; Hymowitz & Freed, 1974). Subsequent studies (Hymowitz, 1976a) showed that the greater sensitivity of induced licking to shock was not simply due to the possibility that shock was more often delivered while the animals were licking than while they were pressing. Schedule-induced licking was similarly suppressed by shock that was separated in time from licking by 1, 5, 10, and 15 sec (Hymowitz, 1976a).

In a recent study, Hymowitz (1977a) employed schedule-induced licking, as well as schedule-controlled lever pressing, as a dependent variable. Food was delivered under a fixed-interval 40-sec schedule. Following the acquisition of stable levels of schedule-induced and schedule-controlled behavior, the animals were exposed on separate occasions to signaled and unsignaled electric-shock delivery. Electric shock was delivered according to a variable-time 70-sec schedule. During signaled shock, a 10-sec signal light preceded each shock. Blocks of successive food-and-shock sessions at a constant shock intensity alternated with blocks of successive food-alone sessions.

The data produced by this study were consistent with the author's previous findings (Hymowitz, 1977b). For both lever pressing and schedule-induced licking, differential response suppression during signaled and unsignaled shock was a function of shock intensity. At some intensities, pressing and licking either were unaffected during signaled and unsignaled shock, suppressed primarily during unsignaled shock, or equally suppressed during signaled and unsignaled shock. Moreover, the intensity of shock required to suppress licking generally was lower than for lever pressing. Hence, licking often ceased during signaled shock delivery, whereas lever pressing was maintained at a high rate. The finding that one behavior occurs at a high rate during signaled shock whereas another is totally suppressed has considerable theoretical importance. It is not clear how such a finding can be explained by Perkins' (1968) preparedness hypothesis or Seligman's (1968) safety signal hypothesis. If the animal is prepared sufficiently to continue pressing or if it is safe to press, why is the animal not prepared sufficiently to continue licking, and why is it not safe to lick?

Differential response suppression during signaled and unsignaled shock also has been reported with operant licking (e.g., Imada & Okamura, 1975). As noted previously, it is important to distinguish between operant and induced licking. In the former, the animals are water deprived and given limited access to water in the test situation. For schedule-induced licking, the animals are not water deprived, and licking occurs as an adjunct (cf. Falk, 1971; Segal, 1972) to ongoing food-

maintained responding. Typically, animals lick after the ingestion of each food pellet (e.g., Falk, 1961). Clearly, Seligman's (1968) original findings pass the test of response generality (cf. Sidman, 1960). However, the available data suggest the response studied may be an important determinant of suppression during signaled and unsignaled shock. Further studies of operant and adjunctive behaviors appear warranted. In particular, it is of considerable interest to determine why schedule-induced licking is suppressed during signaled shock conditions in which schedule-controlled lever pressing is maintained at a high rate. The greater sensitivity of induced licking to shock may simply be due to the nature of the response (lever pressing versus licking), to the relative strength of schedule-induced and controlled behavior per se, or to the rate at which each behavior occurs. With respect to the latter possibility, Blackman (1977) noted that behaviors that occur at a high rate are more affected by a given intensity of shock than are behaviors that occur at a low rate. He referred to this as a *rate-dependency effect*. Since licking generally occurs at a higher rate than does lever pressing, this hypothesis certainly warrants further attention. It is noteworthy, however, that McKearney (1973) showed that the rate of licking was not a factor in determining the effects on induced licking of the drug methamphetamine. With schedule-controlled licking, rate was a major determinant of the drug's effect (rate-dependency effect).

Predictability of Shock

Rescorla (1968) showed that CSs acquired their suppressive or aversive properties only when they reliably specified the advent of shock delivery. When shock occurred with the same probability in the absence and in the presence of the to-be-conditioned stimulus, little conditioning occurred. To what extent is differential response suppression during signaled and unsignaled shock a function of the probability that a CS precedes each shock in the signaled shock condition? Hymowitz (1973a, Experiment 3) employed a 2 × 3 experimental design with sessions as repeated measures. The factors were the schedule of shock delivery (fixed interval or variable interval) and the percentage of occasions on which a 5-sec CS preceded shock delivery —100%, 50%, or 0%.

For the 100% signal condition, the presence and absence of the signal reliably indicated shock and shock-free occasions. For both schedules of shock delivery, the least amount of response suppression occurred in this condition. Responding virtually ceased during the CS, but was maintained at a high rate in its absence. For the 0% signal condition, the presence and absence of shock were not specified. Although the fixed-interval schedule may have provided temporal cues, there was little if any difference in response suppression for the groups of rats exposed to either shock schedule. Both groups revealed marked response suppression with little recovery during the 10 test sessions.

Perhaps the most interesting groups were those for which the CS preceded shock on 50% of the shock occasions. The presence of the CS reliably indicated shock occasions, but the absence of the CS did not reliably indicate shock-free occasions. For variable-interval shock delivery, as much suppression occurred under this condition as under the 0% condition. For fixed-interval shock delivery, in which temporal factors may have specified to some degree the presence and absence of shock, an intermediate degree of response suppression occurred. There was more response suppression than in the 100% condition, but much less than in the 0% condition.

Comparable findings were obtained in preference situations (Badia, Harsh, Coker, & Abbott, 1976; Hymowitz, 1973c). Hymowitz (1973c) used a free-operant shuttle box arrangement in which 10 animals were exposed to a VT 65-sec schedule of scrambled foot shock (.40 mA). For five of the animals, one side of the shuttle box was associated with signaled shock (100%), and the other side was associated with unsignaled shock (0%). For five other animals, one side of the box was associated with unsignaled shock (0%); on the other side, the CS preceded shock on 50% of the shock occasions.

The animals showed a marked preference for 100% signaled shock delivery over 0% or unsignaled shock. This was true of all five of the animals tested. However, a clear preference for the 50% signal over the 0% signal condition was not found. A signal that reliably indicated the presence of shock but not the absence of shock was not preferred over the no-signal condition.

Badia et al. (1976) used a changeover procedure (from unsignaled to signaled shock delivery) to answer a related question. Is changeover responding controlled by stimuli that reliably indicate the presence of shock or by stimuli that reliably indicate the absence of shock? When the CS precedes shock 100% of the time, the CS necessarily serves both signaling functions. When the probability of the CS preceding shock is varied, the signal function of the CS also is altered, depending on the manner in which the CS and the shock are programmed.

In one condition, all of the shocks were preceded by a CS. However, the probability of a CS being followed by shock varied from 1.0 to .02. Thus, at some CS–shock probabilities, the presence of the CS did not reliably indicate shock occasions, although the absence of the CS reliably indicated shock-free occasions. In this condition, the animals consistently changed over from unsignaled to signaled shock at each probability level; that is, animals readily selected the condition in which the CS dependably specified the absence of shock but not necessarily the presence of shock.

In a second condition, all of the CSs were followed by shock. Some shocks were not preceded by a CS, depending on the desired probability value. This is the same manner of programming the CS and the shock that was used by Hymowitz (1973a, 1973c). As the dependability of the CS as an indicator of the absence of shock decreased, the animals' preference for the CS condition decreased, even though the CS still reliably specified the presence of shock. These data are in close agreement with the preference data reported by Hymowitz (1973c) and with the suppression data for the group of animals exposed to the variable-interval shock delivery schedule (Hymowitz, 1973a, Experiment 3).

One other study also examined differential associations between the CS and shock (Nageishi & Imada, 1974). They presented an average of three shocks per session. For one group of animals, all of the shocks were preceded by a CS (100%); for another, two of the shocks were preceded by a CS (66%); for another, one of the shocks was preceded by a CS (33%); in a final group, none of the shocks were preceded by a CS (0%).

Nageishi and Imada (1974) reported that the degree of response suppression was directly related to the degree of shock predictability, the least suppression occurring in the 100% condition and the most in the 0% condition. In their study, the presence of the CS reliability indicated the availability of shock in the 66% and 33% conditions. However, the absence of the CS did not reliably indicate the absence of shock. Based on the preference data cited earlier (Badia et al., 1976; Hymowitz, 1973c), Nageishi and Imada's (1974) findings are somewhat surprising. The preference data suggest that the utility of the CS derives from its specification of shock-free, not shock, occasions.

Nageishi and Imada's (1974) findings are in close agreement with Hymowitz's (1973a) findings for fixed-interval shock delivery but not for variable-interval shock delivery. Although procedural and methodological differences between the studies prohibit direct comparisons, one cannot help but wonder whether the small number of shocks (three per session) delivered in Nageishi and Imada's study in some way accounts for some of the differences. For example, the delivery of the third shock may have signaled shock-free occasions.

Clearly, the percentage of occasions on which the CS precedes shock is an important determinant of response suppression during signaled and unsignaled shock delivery. Moreover, the manner in which the CS and shock are programmed is also an important factor. To gain a better understanding of the relationship between the predictability of shock and differential response suppression, further studies are required. One useful approach would be to investigate the differential sup-

pression of schedule-controlled and schedule-induced responding, perhaps within the same animal, under conditions in which the percentage of occasions on which the CS precedes shock is varied. The CS and shock also should be programmed so that the CS may dependably specify shock-free occasions or dependably specify shock occasions, but not both (cf. Badia et al., 1976).

Discriminability of Shock-Free Occasions

There are two important theoretical viewpoints from which to view data on signaled and unsignaled shock. According to Perkins (1968), the importance of the CS derives from the fact that it enables the animal to prepare for shock. According to Seligman (1968; see also Seligman, Maier, & Solomon, 1971), the CS is important because it specifies occasions in which shock is not available. Such shock-free occasions may allow the animal to relax (cf. Denny, 1971) instead of remaining in a chronic state of fear. The reinforcing effects of shock-free occasions have been demonstrated quite eloquently in studies of avoidance and shock-frequency reduction (Hineline, 1977).

Recent studies by Hymowitz (1976b, 1977b), in which animals were studied under mixed as well as multiple schedules of signaled and unsignaled shock, provide considerable support for the safety signal analysis. Multiple schedules contain discriminative stimuli that distinguish one component from the other. Mixed schedules do not. During multiple schedules of signaled and unsignaled electric-shock delivery, the animals may readily discriminate that component in which the absence of the CS is associated with the absence of shock. During mixed schedules, this discrimination is highly unlikely. Although both schedules are the same in that the CS is reliably associated with the presence of shock, they differ in that only during the multiple schedule is the absence of the CS clearly associated with the absence of shock.

To the extent that the preparedness value of the CS is the determining factor, one would predict differential response suppression during signaled and unsignaled shock for each schedule. To the extent that the discrimination of shock-free time is necessary for differential responding, differential response suppression is predicted for the multiple schedule but not for the mixed schedule.

The data (Hymowitz, 1976b, 1977b) clearly favor the safety signal hypothesis. Although differential suppression was found for the multiple schedule of signaled and unsignaled shock, responding during the components of the mixed schedule of signaled and unsignaled shock was not differentially suppressed. Responding was similarly suppressed during signaled and unsignaled shock delivery. These findings, as well as those of Hymowitz (1973a, 1973c) and Badia et al. (1976) referred to previously, strongly suggest that the discrimination of shock-free occasions, rather than shock occasions, is one of the key determinants of differential responding during signaled and unsignaled shock (see also Arabian & Desiderato, 1975).

Discussion

As noted previously, studies of response suppression during signaled and unsignaled electric-shock delivery uniformly support Seligman's (1968) original report of more response suppression during unsignaled than during signaled shock. Moreover, the generality of Seligman's (1968) original findings was successfully extended to include operant and adjunctive licking, a variety of shock delivery schedules, a wide range of shock parameters, and several different schedules of food delivery. In general, the findings on differential response suppression also are consistent with Seligman's (1968) safety signal hypothesis. In particular, studies that show that the predictability of shock-free occasions, not shock occasions, is a key determinant of differential responding (Hymowitz, 1976b, 1977a) lend strong support to the safety signal hypothesis.

It is also important to note that it is highly unlikely that the findings on differential response suppression are, in some way, due to procedural or methodological artifacts. All of the studies mentioned in this review employed scrambled electric-shock delivery. Hence, modification of shock by postural adjustment was minimized. Some of the studies

(Hymowitz, 1973a, 1976b) used response-dependent shock, and the response lever was included in the shock circuit. With such a procedure, it was nearly impossible for the animal to escape shock once it was produced. Yet, much more response suppression was found during unsignaled than during signaled shock. In addition, Hymowitz (1976b, 1977b) showed that responding was suppressed as much during the signaled as during the unsignaled shock component of mixed schedules of shock delivery. If the animal had used the preshock stimulus as a discriminative cue for an avoidance–escape response, one would have expected less suppression in the signaled component of the mixed schedule. To the contrary, the animals seemed not to benefit from signaled shock delivery unless the absence of the signal was reliably associated with the absence of shock. Such findings are highly compatible with the safety signal hypothesis.

The literature review also revealed a number of variables that seem to influence differential response suppression during signaled and unsignaled shock and choice of signaled over unsignaled shock in the same manner. Both depend on the discriminability of shock-free occasions and the intensity of shock. The agreement between studies of choice and studies of response suppression is so great in some instances that the confidence one may place in the reliability of the data is markedly enhanced. Despite the serious criticisms leveled at much of the choice literature, the good fit between the choice and suppression literature enhances our confidence in the findings that animals do prefer signaled over unsignaled shock and that the variables that control the choice of signaled shock also determine the differential suppression of responding during signaled and unsignaled shock.

Although the data generated during the past decade are generally consistent with the safety signal hypothesis, some of the more recent findings suggest the need for possible revision or extension of the hypothesis. It is not clear, for example, why differential response suppression should occur for any given animal at one intensity of shock or for one kind of behavior and not at another intensity or for another behavior. The discriminability of shock-free occasions plays an important role in determining differential responding, but it is not the only variable of importance. As in most other analyses of response suppression, the resistance of responding to the suppressive effect of shock depends on the interplay of a host of food-related and shock-related behaviors. As additional analyses of the variables that influence differential response suppression become available, it may be possible to fit the safety signal hypothesis within a broader theoretical context. *Safety* and *danger* are relative terms, lacking in precision and clarity. As additional data emerge, it may be more fruitful to analyze the suppression of responding during signaled and unsignaled shock in terms of interactions among the signaling of shock, the motivational state of the organism, the nature of the food and shock schedules, the strength of the response under study, and the past experimental history of the organism. Viewed in this manner, safety and danger are but two of the many variables that determine the suppression of responding. Thus, the time may be rapidly approaching when the safety signal hypothesis will lose its appeal as an explanatory mechanism. However, as shown in the present literature review, the safety signal hypothesis has served us well. It has proved to be a useful organizing principle for a considerable body of experimental literature and has laid the foundation for our current knowledge of the effects of signaled and unsignaled aversive events on behavior.

References

Arabian, J. M., & Desiderato, O. Preference for signaled shock: A test of two hypotheses. *Animal Learning & Behavior*, 1975, *3*, 191–195.

Azrin, N. H. Some effects of two intermittent schedules of immediate and nonimmediate punishment. *Journal of the Experimental Analysis of Behavior*, 1956, *1*, 183–200.

Azrin, N. H., & Holz, W. C. Punishment. In W. K. Honig (Ed.), *Operant behavior: Areas of research and application*. New York: Appleton-Century-Crofts, 1966.

Badia, P., & Culbertson, S. The relative aversiveness of signalled versus unsignalled escapable and inescapable shock. *Journal of the Experimental Analysis of Behavior*, 1972, *17*, 463–471.

Badia, P., Culbertson, S., & Harsh, J. Choice of longer or stronger signalled shock over shorter or weaker unsignalled shock. *Journal of the Experimental Analysis of Behavior,* 1973, *19,* 25–32.

Badia, P., & Harsh, J. Further comments concerning preference for signaled shock conditions. *Bulletin of the Psychonomic Society,* 1977, *10,* 17–20. (a)

Badia, P., & Harsh, J. Preference for signaled over unsignaled shock schedules: A reply to Furedy and Biederman. *Bulletin of the Psychonomic Society,* 1977, *10,* 13–16. (b)

Badia, P., Harsh, J., Coker, C. C., & Abbott, B. Choice and the dependability of stimuli that predict shock and safety. *Journal of the Experimental Analysis of Behavior,* 1976, *26,* 95–111.

Berk, A. M., Marlin, N. A., & Miller, R. R. Systems for delivering tailshock to freely ambulatory rats. *Physiology & Behavior,* 1977, *19,* 815–818.

Biederman, G. B., & Furedy, J. J. Preference-for-signaled-shock phenomenon: Effects of shock modifiability and light reinforcement. *Journal of Experimental Psychology,* 1973, *100,* 380–386.

Biederman, G. B., & Furedy, J. J. Preference for signaled shock in rats? Instrumentation and methodological errors in the archival literature. *The Psychological Record,* 1976, *26,* 501–514. (a)

Biederman, G. B., & Furedy, J. J. The preference-for-signaled-shock phenomenon: Fifty days with scrambled shock in the shuttlebox. *Bulletin of the Psychonomic Society,* 1976, *7,* 129–132. (b)

Blackman, D. E. Conditioned suppression and the effects of classical conditioning on operant behavior. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Bolles, R. C. Reinforcement, expectancy, and learning. *Psychological Review,* 1972, *79,* 394–409.

Bond, N. W., Blackman, D. E., & Scruton, P. Suppression of operant behavior and schedule-induced licking in rats. *Journal of the Experimental Analysis of Behavior,* 1973, *20,* 375–383.

Camp, D. S., Raymond, G. A., & Church, R. M. Response suppression as a function of the schedule of punishment. *Psychonomic Science,* 1966, *5,* 23–24.

Camp, D. S., Raymond, G. A., & Church, R. M. The temporal relationship between response and punishment. *Journal of Experimental Psychology,* 1967, *74,* 114–123.

Church, R. M. Response suppression. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior.* New York: Appleton-Century-Crofts, 1969.

Church, R. M., Raymond, G. A., & Beauchamp, R. D. Response suppression as a function of intensity and duration of a punishment. *Journal of Comparative and Physiological Psychology,* 1967, *63,* 39–44.

Church, R. M., Wooten, C. L., & Matthews, T. J. Discriminative punishment and the conditioned emotional response. *Learning and Motivation,* 1970, *1,* 1–17.

Crabtree, M. S., & Kruger, B. M. Free choice of signaled vs. unsignaled scrambled electric shock with rats. *Bulletin of the Psychonomic Society,* 1975, *6,* 352–354.

D'Amato, M. R. *Experimental psychology.* New York: McGraw-Hill, 1970.

Davis, H., Memmott, J., & Hurwitz, H. M. B. Effects of signals preceding and following shock on baseline responding during a conditioned-suppression procedure. *Journal of the Experimental Anlysis of Behavior,* 1976, *25,* 263–277.

Denny, M. R. Relaxation theory and experiments. In F. R. Brush (Ed.), *Aversive conditioning and learning.* New York: Academic Press, 1971.

Dunham, P. J. Punishment: Method and theory. *Psychological Review,* 1971, *78,* 58–70.

Estes, W. K. Outline of a theory of punishment. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior.* New York: Appleton-Century-Crofts, 1969.

Falk, J. L. Production of polydipsia in normal rats by an intermittent food schedule. *Science,* 1961, *133,* 195–196.

Falk, J. L. Control of schedule-induced polydipsia: Type, size, and spacing of meals. *Journal of the Experimental Analysis of Behavior,* 1967, *10,* 199–206.

Falk, J. L. The nature and determinants of adjunctive behavior. *Physiology & Behavior,* 1971, *6,* 577–588.

Ferraro, D. P. Response suppression and recovery under some temporally defined schedules of intermittent punishment. *Journal of Comparative and Physiological Psychology,* 1967, *64,* 133–139.

Freed, E. X., & Hymowitz, N. Effects of schedule, percent body weight, and magnitude of reinforcer on acquisition of schedule-induced polydipsia. *Psychological Reports,* 1972, *31,* 95–101.

Freed, E. X., Hymowitz, N., & Fazzaro, J. A. Effects of response-independent electric shock on schedule-induced alcohol and water intake. *Psychological Reports,* 1974, *36,* 63–71.

French, D., Palestino, D., & Leeb, C. Preference for warning in an unavoidable shock situation: Replication and extension. *Psychological Reports,* 1972, *30,* 72–74.

Furedy, J. J. An integrative progress report on informational control in humans: Some laboratory findings and methodological claims. *Australian Journal of Psychology,* 1975, *27,* 61–83.

Furedy, J. J., & Biederman, G. B. Preference for signaled shock phenomenon: Direct and indirect evidence for modifiability factors in the shuttlebox. *Animal Learning & Behavior,* 1976, *4,* 1–5.

Ghiselin, M. T. *The triumph of the Darwinian method.* Berkeley: University of California Press, 1969.

Harsh, J., & Badia, P. Choice for signalled and unsignalled shock as a function of shock intensity. *Journal of the Experimental Analysis of Behavior,* 1975, *23,* 349–355.

Harsh, J., & Badia, P. H. Temporal parameter influencing choice between signalled and unsignalled

shock schedules. *Journal of the Experimental Analysis of Behavior,* 1976, *25,* 327–333.

Hineline, P. N. Negative reinforcement and avoidance. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Hoffman, H. S., & Fleshler, M. Stimulus aspects of aversive controls: The effects of response-contingent shock. *Journal of the Experimental Analysis of Behavior,* 1965, *8,* 89–96.

Hunt, H. F., & Brady, J. Some effects of punishment and intercurrent "anxiety" on a simple operant. *Journal of Comparative and Physiological Psychology,* 1955, *48,* 305–310.

Hymowitz, N. Comparisons between variable-interval and fixed-interval schedules of electric shock delivery. *Journal of the Experimental Analysis of Behavior,* 1973, *19,* 101–111. (a)

Hymowitz, N. Effects of lever-press-dependent and independent electric shock on schedule-induced water intake. *The Psychological Record,* 1973, *23,* 487–497. (b)

Hymowitz, N. Preference for signaled electric shock. *Proceedings of the 81st Annual Convention of the American Psychological Association,* 1973, *8,* 847–848. (Summary) (c)

Hymowitz, N. Effects of electric-shock delivery on schedule-induced water intake: Delay of shock, shock intensity, and body-weight loss. *Journal of the Experimental Analysis of Behavior,* 1976, *26,* 269–280. (a)

Hymowitz, N. Effects on responding of mixed and multiple schedules of signalled and unsignalled response-dependent electric-shock delivery. *Journal of the Experimental Analysis of Behavior,* 1976, *25,* 321–326. (b)

Hymowitz, N. Effects of signalled and unsignalled electric-shock delivery on schedule-controlled and schedule-induced behavior. *The Psychological Record,* 1977, *27,* 715–731. (a)

Hymowitz, N. Effects of the intensity of electric shock on response suppression during multiple and mixed schedules of signalled and unsignalled electric-shock delivery. *The Psychological Record,* 1977, *27,* 425–440. (b)

Hymowitz, N., & Freed, E. X. Effects of response-dependent and independent electric shock on schedule-induced polydipsia. *Journal of the Experimental Analysis of Behavior,* 1974, *22,* 207–213.

Imada, H., & Okamura, M. Some cues rats can use as predictors of danger and safety. *Animal Learning & Behavior,* 1975, *3,* 221–225.

Jacquet, Y. F. Schedule-induced licking during multiple schedules. *Journal of the Experimental Analysis of Behavior,* 1972, *17,* 413–423.

Kamin, L. J. Temporal and intensity characteristics of the conditioned stimulus. In W. F. Prokasy (Ed.), *Classical conditioning.* New York: Appleton-Century-Crofts, 1965.

Lockard, J. S. Choice of warning signal or no warning signal in an unavoidable shock situation. *Journal of Comparative and Physiological Psychology,* 1963, *3,* 526–530.

MacDonald, L. The relative aversiveness of signalled versus unsignalled shock-punishment. *Journal of the Experimental Analysis of Behavior,* 1973, *20,* 37–46.

MacDonald, L., & Baron, A. A rate measure of the relative aversiveness of signalled vs. unsignalled shock. *Journal of the Experimental Analysis of Behavior,* 1973, *19,* 33–38.

McKearney, J. W. Effects of methamphetamine and chlordiazepoxide on schedule-controlled and adjunctive licking in the rat. *Psychopharmacologia,* 1973, *30,* 375–384.

Millenson, J. R., & de Villiers, P. A. Motivational properties of conditioned anxiety. In R. M. Gilbert & J. R. Millenson (Eds.), *Reinforcement.* New York: Academic Press, 1972.

Miller, R. R., Daniel, D., & Berk, A. M. Successive reversals of a discriminated preference for signaled tailshock. *Animal Learning & Behavior,* 1974, *2,* 271–274.

Miller, R. R., Marlin, N. A., & Berk, A. M. Reliability and sources of control of preference for signaled shock. *Animal Learning & Behavior,* 1977, *5,* 303–308.

Morse, W. H., & Kelleher, R. T. Schedules using noxious stimuli: I. Multiple fixed-ratio and fixed-interval termination of schedule complexes. *Journal of the Experimental Analysis of Behavior,* 1966, *9,* 267–290.

Myers, J. S. Some effects of noncontingent aversive stimulation. In F. R. Brush (Ed.), *Aversive conditioning and learning.* New York: Academic Press, 1971.

Nageishi, Y., & Imada, H. Suppression of licking behavior in rats as a function of predictability of shock and probability of conditioned-stimulus–shock pairings. *Journal of Comparative and Physiological Psychology,* 1974, *87,* 1165–1173.

Orme-Johnson, D. W., & Yarczower, M. Conditioned suppression, punishment, and aversion. *Journal of the Experimental Analysis of Behavior,* 1974, *21,* 57–74.

Perkins, C. C., Jr. An analysis of the concept of reinforcement. *Psychological Review,* 1968, *75,* 155–172.

Perkins, C. C., Jr., Seymann, R. G., Levis, D. J., & Spencer, H., Jr. Factors affecting preference for signal-shock over shock-signal. *Journal of Experimental Psychology,* 1966, *72,* 190–196.

Rachlin, H., & Herrnstein, R. J. Hedonism revisited: On the negative law of effect. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior.* New York: Appleton-Century-Crofts, 1969.

Rescorla, R. A. Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology,* 1968, *66,* 1–5.

Segal, E. F. Induction and the provenance of operants. In R. M. Gilbert & J. R. Millenson (Eds.), *Reinforcement.* New York: Academic Press, 1972.

Seligman, M. E. P. Chronic fear produced by un-

predictable electric shock. *Journal of Comparative and Physiological Psychology,* 1968, *66,* 402–411.
Seligman, M. E. P. On the generality of the laws of learning. *Psychological Review,* 1970, *77,* 406–418.
Seligman, M. E. P., & Binik, Y. M. The safety signal hypothesis. In H. Davis & H. M. B. Hurwitz (Eds.), *Operant-Pavlovian interactions.* New York: Wiley, 1977.
Seligman, M. E. P., Maier, S. F., & Solomon, R. L. Unpredictable and uncontrollable aversive events. In F. R. Brush (Ed.), *Aversive conditioning and learning.* New York: Academic Press, 1971.
Seligman, M. E. P., & Meyer, B. Chronic fear and ulcers in rats as a function of unpredictability of safety. *Journal of Comparative and Physiological Psychology,* 1970, *73,* 202–207.
Sidman, M. *Tactics of scientific research.* New York: Basic Books, 1960.
Staddon, J. E. R. Learning as adaptation. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 2). New York: Erlbaum, 1975.
Staddon, J. E. R. Behavioral competition in conditioning situations: Notes toward a theory of generalization and inhibition. In H. Davis & H. M. B. Hurwitz (Eds.), *Operant-Pavlovian interactions.* New York: Wiley, 1977. (a)
Staddon, J. E. R. Schedule-induced behavior. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1977. (b)
Stein, L., Sidman, M., & Brady, J. V. Some effects of two temporal variables on conditioned suppression. *Journal of the Experimental Analysis of Behavior,* 1958, *1,* 153–162.

Received October 27, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 1, 191–215

Infant Crying as an Elicitor of Parental Behavior: An Examination of Two Models

Ann D. Murray
Macquarie University, Sydney, Australia

Two models of the compelling nature of the infant cry and its effectiveness in eliciting caregiving behavior are examined. The first model is that of the cry as a releaser of parental behavior. It is suggested that a good fit between the available data and this model depends upon broadening the classical definition of the releaser concept to include motivational factors in a manner advocated by some of the modern ethologists. A model of the cry as an activator of motives of an egoistic or altruistic nature is also examined. This model, based on Hoffman's theory of altruistic motivation, contributes to an understanding of both the compelling releaserlike effects of the cry as well as the wide variations observed within and between cultures in the nature and extent of caregiver responsiveness. It is further argued that altruistic behavior toward crying infants in particular must be viewed within the specific context of ontogenetic processes that enhance the attractiveness of the young for adult caregivers.

It is part of the folklore that the cry of a young infant is a compelling stimulus. There is an urgency associated with the cry that makes a response to it obligatory (Ostwald, 1963). In addition to its purely motivational qualities, the cry evokes intense emotional reactions that can be either of a constructive or of a destructive nature. It is capable of evoking strong feelings of concern and protectiveness on the one hand or of extreme hostility on the other (Ostwald, 1963). The action taken in response to the cry can likewise consist of nurturant acts or murderous ones (Stone, Smith, & Murphy, 1973, p. 1002).

In the theoretical and empirical analysis that follows, two conceptualizations of the mechanisms by which the cry has its powerful impact are examined. First, the question of whether the cry can be considered a releaser of parental behavior is discussed with reference to claims for its evolutionary significance, the physical nature of the stimulus, its signal value, parental responses to it, and possible receptor mechanisms. Subsequently, a model is presented that views the cry as a graded signal that activates motives of either an altruisitc or an egoistic nature. Last, some factors influencing the ontogeny of parental behavior are examined. For purposes of this analysis, the focus is restricted to crying in early infancy when the infant is not physically mobile and is totally dependent on a caregiver for the satisfaction of his or her needs.

The Cry as a Releaser of Caregiving Behavior

Evolutionary Significance of the Cry

Within the framework of the attachment theory formulated by Bowlby (1969) and Ainsworth (1969), it has been suggested that the infant cry may serve as a releaser of caregiving behavior. Crying is considered by these theorists to be an attachment behavior that promotes proximity to or contact with the

The preparation of this article was supported by a Commonwealth of Australia Postgraduate Research Award. Thanks are due to Joanne Cornwell, Jacqueline Goodnow, John Murray, and Peter Van Sommers for their comments on earlier drafts of the article.

Requests for reprints should be sent to Ann D. Murray, who is now at Foundation 41, The Women's Hospital, Crown Street, Sydney, N.S.W., 2010, Australia.

caregiver, usually the mother. Using an ethological perspective, these theorists have hypothesized that attachment behaviors such as crying, smiling, and following originally evolved to perform a protective function by bringing the infant into close proximity with the mother, who could then defend him or her against predators and other dangers. Although such close proximity between mother and infant is not necessary to ensure protection in our present society, Bowlby and Ainsworth argued that babies are nevertheless genetically programmed to cry when out of contact or distressed and that their behavior is adapted to the prototype of a responsive caregiver.

To support his thesis concerning the evolutionary adaptedness of close mother–infant proximity and sensitivity to crying, Bowlby cited indirect evidence. He argued that species-typical characteristics are adapted to the environment in which the species evolved. For *Homo sapiens,* it is especially difficult to determine the adaptive significance of characteristics because people have modified their environment so drastically from that in which they originated. The kinds of evidence used to support his theory include comparative studies of animals and anthropological studies of contemporary societies that exist in environments that are least modified from the "environment of evolutionary adaptedness." Since the publication of Bowlby's work in 1969 three recent studies have provided additional support for his theory.

Konner (1972; Devore & Konner, 1974) studied the rearing patterns of a hunter–gatherer society in Botswana. He argued that 98% of the evolutionary history of *Homo sapiens* took place in a hunter–gatherer economy and that very little evolution has occurred since then. He hoped to be able to highlight the adaptive consequences of the rearing pattern he found with reference to the environmental selective pressures associated with this preagricultural nomadic existence. A major feature of the rearing pattern he found was that there was virtually continuous contact between the mother and her infant. Infants were carried in a sling on the mother's side or on the hip. The feeding pattern was also described as continual in that infants were nursed at least twice an hour for a period of 30 sec to 10 min. Konner reported that infants rarely cried because mothers could anticipate hunger by interpreting more subtle proximal cues such as bodily movements, facial expressions, and so on. In fact, the cry was treated as an emergency signal and was responded to as such with an average latency of 6 sec. Note that this contrasts sharply with the practice in Western cultures in which response is delayed for from 5 to 30 min (Bernal, 1972) and young infants cry an average of 1 to $2\frac{3}{4}$ hours a day (Bernal, 1972; Brazelton, 1962). According to Konner, the adaptive consequences of the hunter–gatherer rearing pattern were twofold. Close proximity served as a protection from predators and other dangers in early infancy and fostered strong attachments to adult models of hunting and gathering skills.

Although the comparative evidence presented by Bowlby has been criticized for its heavy reliance on studies using the rhesus monkey (Dolhinow & Olson, Note 1), support has been provided for Bowlby's theory by the findings of Blurton-Jones (1972) in his survey of a wide variety of mammalian species. Blurton-Jones asked whether human infants and their mothers are adapted to relatively continuous contact (the pattern common among hunter–gatherer societies and Old World monkeys and apes) or to relatively discontinuous contact as exemplified by the European and American style of rearing. To answer this question he surveyed a wide variety of mammals to determine what anatomical, physiological, and behavioral features correlated with the two rearing patterns of caching and carrying. He then applied the correlations found in animals to the human mother and infant to predict whether humans evolved as a caching or a carrying species. In caching species such as the tree shrew, the young are left for long periods in hiding places while the mother forages for food, and feeding is at widely spaced intervals. In carrying species, the young ride on the mother, and feeds are closely spaced.

The evidence that human beings evolved as a caching species was not convincing be-

cause (a) the young would have to remain silent until the mother returned so as not to attract predators by crying, (b) cached young do not urinate or defecate without maternal stimulation as this too would attract predators, and (c) the young would have to possess mechanisms of thermoregulation. None of these conditions were satisfied for the human infant.

The evidence for carrying, however, strongly favored continuous contact. Three arguments were advanced based on milk composition, sucking frequency, and the duration of feeding. The composition of the mother's milk correlated with the schedule of feeding such that in widely spaced feeders, there was a high protein and fat content, whereas in continuous feeders, there was a low protein and fat content. For example, the tree shrew, which feeds every 48 hours, had a high protein and fat content compared with the higher nonhuman primates, which are continuous feeders. Human milk was found to be identical in fat and protein content to that of the continuous-feeding anthropoid apes. Similarly, sucking frequency correlated with feeding frequency and milk composition. Animals that fed the least often sucked the fastest, and fast sucking was associated with a high protein and fat content in the mother's milk. The slow sucking rate in human infants is, therefore, adapted to a pattern of continuous contact. The third piece of evidence was that widely spaced feeders fed for a short duration. For example, the duration of feeding for rabbits was 4 to 5 minutes once every 24 hours. The time spent each day in feeding for human infants was found to be comparatively long, suggesting adaptation to the continuous-feeding pattern.

As further support for the hypothesis that human infants are adapted to the prototype of a responsive mother, Bell and Ainsworth (1972) have claimed that prompt maternal responses to crying early in the first year have adaptive outcomes for infants. They found that responsiveness to crying in the first half year led to a decrease in the frequency and duration of crying in the second half year. In addition, those infants whose mothers had responded promptly to their cries early on were less likely to use crying in an instrumental manner at 1 year of age and were more likely to develop other social signals such as bodily gestures, facial expressions, and vocalizations to communicate with their mothers. According to Bell and Ainsworth, their findings indicate that crying is at first expressive and indiscriminate and that only toward the end of the first year does crying become "goal corrected" and used with the intent to influence others.

If the infant cry is an adaptive species-characteristic behavior that evolved as an emergency signal, one might expect to find reciprocal mechanisms in caregivers that ensure a response in kind to the cry. The consequences of an appropriate response for the infant relate to his or her chances of survival; an appropriate response is adaptive for the parent also because the infant's survival ultimately contributes to the parent's reproductive success. In cases such as this, in which contact between conspecifics is adaptive to both, the sender of the signal and the receiver of the signal will be mutually adapted to each other (Eibl-Eibesfeldt, 1975). In this vein, it has been suggested that the infant cry and other attachment behaviors such as smiling may operate as sign stimuli that release parental behavior (Bowlby, 1958).

The Releaser Mechanism

Since Lorenz's original observations of releasers in the 1930s, examples of this form of adaptive mechanism have been described in many species (Eibl-Eibesfeldt, 1975). A sign stimulus is a simple, conspicuous, and specific stimulus that acts figuratively as the key that unlocks a stereotyped motor action (often referred to as a fixed action pattern). The receiver of the key stimulus is said to possess an innate releasing mechanism (IRM), the hypothesized neural receptor center for receiving and filtering afferent impulses produced by the key stimulus. Releasers have been found to regulate interactions between prey and predator, between sexual partners, between sexual rivals, and between parent and young in many species (Eibl-Eibesfeldt, 1975).

Although the infant cry is often casually likened to a releaser, no formal assessment of whether it fits the model has ever been carried out. Much effort has been put into describing the physical characteristics of the stimulus, but little effort has been made to determine whether and/or which key features affect parental responses. Often the question of whether the cry acts as a releaser is linked with the issue of whether different cry types, for example, for pain versus hunger, are recognizable and how these cry types differ in physical characteristics. In the ensuing paragraphs, an attempt is made to assess the "goodness of fit" of the model of the cry as a releaser.

Although it is recognized that the releaser concept as originally proposed by Lorenz (1937) has recently come under attack (cf. Hinde, 1974; Klopfer, 1974), the classical view of the IRM has been adopted here for heuristic purposes. The following list of the basic characteristics of releaser systems is derived from Eibl-Eibesfeldt (1975), who provides the modern expression of the Lorenzian ethological view.

1. The basis for recognition of the key features of the stimulus and for the performance of a response is said to be *innate* in that the system is functional in inexperienced animals that have had no prior experience with the stimulus.

2. Sign stimuli are made up of simple, conspicuous, and specific cues. It is even possible to trick an animal into responding to simplified (artificial) models of the stimulus in an unnaturalistic context.

3. Sign stimuli exhibit the phenomenon of *heterogeneous summation*. The same response can be elicited by several different and independent cues that combine additively to increase the effectiveness of the sign stimulus.

4. The optimal stimulus is often unrealistic. By exaggerating key features of the stimulus, a "supernormal" stimulus is produced that is more effective than a natural stimulus.

5. The response to the stimulus is a stereotyped behavioral response.

6. This stimulus–response specificity is accomplished through a neural receptor–effector system in which afferent impulses from the senses are linked with efferent impulses from motor centers leading to the occurrence of fixed action patterns.

Each of these characteristics is reviewed in relation to the available literature on infant crying.

Innateness of Stimulus Recognition

For all practical purposes, the criterion of *innateness* cannot be examined for adults who, even if never exposed to an infant's cry, cried as infants themselves and have had the experience of crying during their life spans. The only research that possibly bears on this issue are the studies of contagious crying in newborns conducted by Simner (1971) and Sagi and Hoffman (1976). Simner found that infants tested at about the age of 70 hours cried more in response to a tape recording of a newborn cry than they did to either the cry of a 5½-month-old infant or to two artificially produced sounds (a computer-synthesized cry designed to contain features similar to the newborn cry and a series of white noise bursts included to control for the nonvocal properties of the cry). Furthermore, a tape recording of the infant's own cry produced more crying than the cry of another newborn. Contagious crying was found to be greater among female than among male infants. Simner speculated that imitation of sounds not very discrepant from those with which the infant was already familiar could account for the greater effectiveness of the two newborn cries when compared with the cry of the older infant and with the artificial sounds.

Sagi and Hoffman (1976) replicated Simner's (1971) results with younger infants (average age was 34 hours) using only the newborn cry and the synthesized cry from Simner's tapes. The newborn cry elicited more crying than did the synthetic cry and a control period of silence. The sex difference in contagious crying was also replicated. The authors took issue with Simner's original cognitive interpretation of these findings, because the contagious crying they observed appeared to be indicative of genuine distress rather than just a vocal response in imitation of a vocal stimulus. They interpreted their

results in terms of inference theory: Distress cues from another person evoke associations with the observer's own past distress and result in a distressed state in the observer. This interpretation rested on a classical-conditioning paradigm. Because they felt that evidence for conditioning in 1-day-old infants was weak, Sagi and Hoffman suggested that the distress response to another's distress could be innate.

The Cry as a Simple, Conspicuous Stimulus

Studies of cry sounds have been not only numerous but also inspired by many diverse interests on the part of investigators. These studies fall into two groupings: (a) detailed acoustical descriptions of cries and cry types inspired by the new technology of the sound spectrograph and other instruments (normative studies) and (b) research on the ability to recognize cry types (signal value studies).

Normative studies. Although there was one early attempt at an acoustical study of the infant cry (Fairbanks, 1942), most of the normative studies followed the invention of the spectrograph. These normative studies have established the temporal pattern of the newborn cry; it can be described as an expiratory cry that lasts .6 to 1.4 sec and is followed by a brief silence (.2 sec); it is sometimes followed by a brief inspiratory whistle (.1 to .2 sec) and by another brief period of rest (.2 sec) before another expiratory cry begins (Sedlackova, 1964; Truby & Lind, 1965; Wolff, 1969). The fundamental frequency of the expiratory cry is 400 Hz on the average, with the inspiratory whistle having a somewhat higher (550–600 Hz) frequency (Truby & Lind, 1965; Wolff, 1969). There is often a rising–falling pattern of frequency, for example, an expiratory cry of .6 sec in length might begin with a fundamental frequency of 300–350 Hz for .2 sec, rise to 500 Hz for .2 sec, and fall to 300–200 Hz for .2 sec (Truby & Lind, 1965; Wolff, 1969). The loudness of the cry is about 80 dB when measured 12 inches (30.5 cm) from the baby's mouth (Ringel & Kluppel, 1964). There are wide individual differences among babies in the characteristics of their cries, but individual cry patterns have been found to be stable for each infant (Ringel & Kluppel, 1964).

Before the emphasis in cry studies shifted to the description of cry types associated with different causes or emotions, Truby and Lind (1965) distinguished three general types of expiratory cries based on their inspection of hundreds of sound spectrograms. These types appeared to be associated with the intensity or effort put into the cry. The cries they analyzed were obtained from 1- to 12-day-old infants who were all pinched to provide a standard stimulus to cry. The three cry types were phonation, dysphonation, and hyperphonation. They called the basic cry pattern phonation because of its harmonic structure and the symmetry and smoothness of the spectrogram and intensity pattern. Phonated cries did not give the impression of great distress or discomfort. Dysphonated cries were effortful performances in which turbulence or noise caused by overloading at the larynx obscured the harmonics of the basic cry pattern. Hyperphonated cries were characterized by an abrupt shift from or to a very high pitch of up to 2000 Hz, were whistle-like, and were caused by strain and constriction of the vocal apparatus. Hyperphonation often occurred concurrently with dysphonation for the most vociferous response to discomfort. Strain during egression was usually followed by hyperphonation during ingression—the inspiratory whistle described earlier.

The major attempts to differentiate acoustically among cry types of newborns according to their cause or underlying emotion have been made by Wolff (1969) and Wasz-Hockert, Lind, Vuorenkoski, Partenen, and Valanne (1968). Wolff distinguished three major cry types: the hunger cry, the mad or angry cry, and the pain cry. The hunger cry had no causal relation with hunger according to Wolff but was really just a basic rhythmical pattern. The basic pattern that Wolff described is similar to the phonated cry in Truby and Lind's classification. The inspiratory whistle that Wolff found to occur within the basic cry pattern can be likened to the hyperphonated cries of the latter in-

vestigators. Wolff's mad or angry cry was characterized by turbulence due to an excess of air being forced through the vocal cords. This mad or angry cry was, then, synonymous with the dysphonated cry described by Truby and Lind. Wolff's third cry type, the pain cry, was differentiated not on the basis of acoustical attributes, but on the basis of the temporal pattern of the first two or three expiratory cries following application of a painful stimulus. Wolff's pain cries were recorded during the general hospital practice of pricking the neonate's heel to take a blood sample. This painful stimulus probably produced more discomfort than the pinch used by Truby and Lind. The cardinal features of Wolff's pain cry were (a) a sudden onset of loud crying (as opposed to a gradual buildup for the hunger or basic cry), (b) an initial long cry (as long as 4 sec compared with 1 sec or less for the basic cry), and (c) an extended period of breath holding after the initial cry (for as long as 7 sec). In Wolff's study, the pain cry settled down to the basic temporal pattern after two or three long expiratory cries. Except for pointing out the temporal pattern of the cry following a very painful experience, Wolff's formulation adds little to that provided by Truby and Lind. Wolff himself disavowed that the basic cry had a causal connection with hunger, and the mad or angry cry can be described as just a more effortful performance indicative of greater discomfort, as in Truby and Lind's formulation. For convenience in the following discussion of research concerning cry types, the labels used by the authors concerned, for example, "hunger" cry, have been retained, although the implied causal connections may be questionable.

Following what appeared to be a rather simple and appealing description of the basic cry pattern (phonation) with various superimposed features indicative of effort (dysphonation, hyperphonation, and length of the expiratory cry), Wasz-Hockert et al. (1968) performed a multivariate study of characteristics of cry types based on 11 attributes: length of the expiratory cry, pitch (minimum, general, and maximum), shift (cf. hyperphonation), voice (voiced, voiceless, and half voiced; cf. phonation vs. dysphonation), melody types (rising–falling, rising, falling, flat, and no melody), continuity of signal, glottal plosives, vocal fry, nasality, tenseness, and subharmonic break. Using examples of birth, hunger, and pain cries, a multiple discriminant function analysis was used to arrive at decision rules for classifying cries into types. A rising–falling melody was associated with hunger cries; any melody other than rising–falling and with a length of more than 1.5 sec was a pain cry, and melodies other than rising–falling with lengths of less than 1.5 sec were birth cries. Although shift (hyperphonation) and voice (phonation vs. dysphonation) did not contribute substantially to the classification based on the multiple discriminant function analysis, the authors mentioned that shift occurred most frequently with pain cries, that birth cries were usually voiceless, and that hunger and pain cries did not differ in voice. Given that so few of the measures used were predictive and that the major bases for differentiation could be said to be the length of the signal and types of phonation (because the scoring of melody was confounded with voice, i.e., voiceless sounds could not be scored for melody), the Wasz-Hockert results seem to indicate that the cries were not uniquely different according to what caused them but rather differed in intensity according to the degree of discomfort experienced by the infant. One might expect that a baby experiencing hunger will be less distressed than one experiencing birth or pain. The authors themselves noted that hunger cries when left a long time unattended began to resemble pain cries.

Signal value studies. The studies conducted to determine whether different cry types are recognizable are fraught with methodological differences that make conclusions difficult to reach. Some studies have used multiple-choice techniques, others have not. Some have controlled for the durations of the cry sequences, others have not. One has used single expiratory cries, thus eliminating information provided by the temporal pattern of cries. The choices of cry types to be identified and the methods of evoking the cries have differed. In some studies, investigators,

have narrowed their selection of cries only to cries that they say are "typical" of their type.

The landmark study of this type was conducted by Sherman (1927). Without the benefit of magnetic recording devices, he had to make live presentations of his stimuli. Groups of observers separated from the infants by a screen listened to cries elicited by four stimuli: hunger, sudden dropping, restraint of the head and face, and sticking with a needle. Observers were exposed to each cry for approximately 10 to 15 sec. With no predetermined cry-evoking categories provided, the 23 observers exhibited little agreement and mentioned a total of 12 emotions including hunger, pain, fear, colic, rage, and so forth. Sherman noted that when observers were allowed to listen for a longer period of time (2 to 3 min), they were able to distinguish between cries caused by the application of an external stimulus and cries associated with an internal organic condition (hunger or colic), as the latter tended to be more prolonged than the former.

The strongest criticisms of Sherman's work have been voiced by Izard (1971) and Ekman (Ekman, Friesen, & Ellsworth, 1972). According to Izard, the most remarkable thing about Sherman's work is that no one questioned why observers should be expected to differentiate among emotional reactions that all included distressful crying. Izard considers crying to be primarily the expression of one emotion: distress–anguish. This emotion is equivalent to what Darwin (1872/1965) described as the emotion of suffering. Ekman criticized Sherman for failing to distinguish between judgments of emotions and judgments of events and for treating responses of rage and anger as separate and pain and hurt as separate when these are synonyms. Further, he echoed Izard's concern that all four situations may have elicited the same emotion. Whereas Ekman suggested that the predominant emotion expressed may have been anger–rage based on the frequency with which observers nominated this emotion, Izard argued that distress would have predominated over angry emotions.

The next series of studies was conducted by Wasz-Hockert and his colleagues. In the first study (Wasz-Hockert, Partenen, Vuorenkoski, Michelsson, & Valanne, 1964), nurses listened to six examples of "typical" cries of four types: birth, hunger, pain, and pleasure. The length of the cry samples varied from 5 to 17 sec with a mean length of 12.3 sec. Nurses correctly identified an average of 16 out of the 24 cries using a multiple-choice method in which they were told to select from among the four cry types. The poorest score of 11 was significantly different from a chance score of 6.

In the next study (Wasz-Hockert, Partenen, Vuorenkoski, Valanne, & Michelsson, 1964), these investigators used the same method to test males and females who were experienced and inexperienced caregivers. Experienced women included mothers, children's nurses, pediatricians, and midwives. Experienced men included fathers and pediatricians. Inexperienced men and women had not looked after infants from 0 to 2 years in age for as long as 2 weeks. Experienced women correctly identified significantly more cries than inexperienced women. Experienced men did not significantly differ from inexperienced men. The lowest group mean was still significantly different from a chance level. Although the authors did not test for a sex difference, their data indicate that men were considerbly less accurate than women. Surprisingly, fathers and pediatricians scored lower than women with no experience. In addition, the easiest cry to identify was the pleasure cry, which was identified correctly almost 100% of the time. The hardest to identify was the birth cry.

In a later study with only women as subjects (Wasz-Hockert et al., 1968), a slightly different method was used. Six examples of each of the four cry types were randomly selected from the pool of expiratory cries used in the multiple discriminant function analysis described earlier. Each single expiratory cry was repeated seven times. Again, experienced women were found to perform more accurately than inexperienced women, and the lowest group score was higher than chance level. Again, the pleasure cry was easiest to identify and the birth cry the hardest. Birth was often confused with hunger and pain, and

hunger and pain were often confused with one another. The most poorly recognized cries, the authors claimed, were not representative of their type.

Muller, Hollien, and Murry (1974) criticized the Wasz-Hockert studies on a number of grounds. In particular, they felt that the inclusion of birth and pleasure cries was inappropriate and that Wasz-Hockert should have controlled for sample duration. They reported evidence that subjects could not correctly identify cries evoked by pain stimulation (snapping a rubber band against the skin), auditory stimulation (clapping together wooden blocks), and hunger stimulation (withdrawal of feeding). Note that this last cry was evoked differently from previous studies in which naturally occurring hunger cries were recorded before a meal. Mothers of the infants recorded, and a group of mothers of infants of comparable age listened to, two 15-sec segments of crying for each stimulus. Neither group of mothers was able to identify correctly the cry-evoking situation in a multiple-choice situation. Muller et al. concluded that the acoustic characteristics of cries carry little perceptual information about the cry-evoking situation, that the cry generally acts only to alert the mother, and that judgments of the causes of crying in the home are based on additional environmental cues.

Interest in whether cry types are identifiable has stemmed from two sources. This issue is of importance not only to those who have been concerned with the ability to correctly identify emotional expressions of others but also to those who have been tracing the continuity of development of infant sounds for communication from crying to babbling and finally to language (e.g., Irwin, 1948; Lynip, 1951; Winitz, 1960). It is possible that this intense concern represents an overintellectualization of the problem. On the basis of the confusion in the literature over different cry types, their unique characteristics, and their identifiability, it is more likely that certain attributes relate to the intensity of discomfort or distress felt and that any accuracy of identification according to evoking situation is due to a correlation between the intensity of the cries and observers' prior notions about the intensity of negative emotion associated with particular cry-evoking situations. Hunger builds up slowly and is ordinarily not as distressing an experience as pain for adults; therefore, raucous cries with sudden onsets are classified as due to pain and melodious cries with slow buildups as due to hunger. Indirect support for this hypothesis could be derived from the fact that there is no agreement in the literature on the causes of crying (see Aldrich, Sung, & Knop, 1945a; Aldrich, Sung, & Knop, 1945b; Brazelton, 1962; Illingsworth, 1955; Lakin, 1957; Lennane & Lennane, 1973; Stewart et al., 1954; and Wolff, 1969 for lists and discussions of possible causes of crying), and therefore one would not expect there to be agreement on the cry types associated with different causes.

Attempts to read into the cry more than is there may be partly a by-product of the distance maintained between parent and infant in our culture. Because this distance is so great as to preclude the use of more subtle communications (e.g., body movement and facial expression), great effort has gone into interpreting the only distal signal available to the infant—the cry. And further, because psychological needs (e.g., for contact or cuddling) as opposed to physiological needs are often not considered legitimate and deserving of attention, there is a need on the part of the parent to differentiate between "real" cries and "fake" cries. Reading into cries differentiated signal values with a view to elucidating the beginnings of prespeech communicative competence also belies the discontinuity between crying and the development of communicative competence. For example, Bell and Ainsworth (1972) found that excessive crying was negatively related to the development of positive social communicatory skills. This point is further amplified in a later section when evidence that there are two distinct systems controlling the expression of voluntary and involuntary vocalizations in human beings is reviewed.

The question I have addressed in the present section is whether the cry is a simple, conspicuous stimulus with key features important for stimulus recognition. It can be argued that the key features that emerge in

this review of the extensive literature relate more to stimulus intensity than to stimulus recognition. Whether there are simple features that are essential for recognition would have to be determined by systematically varying the acoustic parameters of the cry and noting which cues are essential for identification of the sound as an infant cry. In results mentioned earlier, infants differentiated between a real newborn cry and a synthesized cry (Sagi & Hoffman, 1976; Simner, 1971). This may indicate, by the criterion that the organism can be tricked into responding to a simple model of a releaser, either that the cry is not a releaser or that the essential key features were overlooked in synthesizing the cry. On the other hand, casual observation provides evidence that adults are in fact frequently tricked by the cry of a Siamese cat even when it is out of context, for example, when no infant is expected to be in the vicinity of the listener, who is not even a highly motivated parent.

Principle of Heterogeneous Summation

According to the principle of heterogeneous summation, separate and noninteracting cues combine additively to produce stimulus recognition. For example, the male stickleback's fighting response to a rival male is elicited either by the red spot on the male's belly or by the head-down position of the rival male (Eibl-Eibesfeldt, 1975). However, when both cues are combined, the attack response is more reliably elicited, that is, the stimulus is more likely to be recognized.

It might be argued that there are separate key features that, when superimposed on the basic cry pattern, would be more likely to elicit parental responses. These key features might include dysphonation or turbulence, hyperphonation or a shift to high pitch, the length of the expiratory cry, and the suddenness of onset. In a small experiment, Wolff (1969), while conducting his naturalistic observations of infants, played a tape recording of the infant's basic hunger cry and, on a separate occasion, played a pain cry to measure the delay before the mother responded. He claimed that there was a dramatic difference in speed of response, with response to the pain cry being almost immediate. As the pain cry was likely to possess some of the markers of intensity mentioned above, Wolff's result may provide some evidence of heterogeneous summation.

Evidence for Supernormal Stimulation

Although no one has tried to produce a supernormal cry stimulus artificially to measure its effectiveness, studies of the cries of brain-damaged and other abnormal infants may be of relevance to this issue. During the past 10 years, Wasz-Hockert and his colleagues have compared the cries of normal infants with those of abnormal infants, including those with Down's syndrome, neonatal asphyxia, brain damage, hyperbilirubinemia, *cri du chat* syndrome, and mixed or unspecified pathology. These studies, summarized in two recent reviews (Vuorenkoski, Lind, Wasz-Hockert, & Partenen, 1971; Wasz-Hockert et al., 1968), have found that the cries of abnormal babies are higher or lower in pitch than normal cries, have greater variability, and have different temporal patterns marked by either shorter or longer cry durations and cry intervals. The cries of abnormal infants are said to be so unpleasant that they override differences in maternal style (Ostwald, 1973; Wolff, 1969). Nurses expend much energy to keep them quiet, or they stay out of earshot. The cribs of these babies are said to be often tucked away in the farthest corner of the nursery from the nursing station because the cries of these infants are so unbearable to listeners (Milowe & Lourie, 1964). The exaggerated acoustic features and temporal patterns of the cries of abnormal babies may constitute supernormal stimulation, but no research has been reported that systematically relates these features to the effect of these cry sounds on the listener.

Evidence for a Stereotyped Response

The question of whether the cry produces a stereotyped response can be dealt with in two parts: First, does the cry elicit a response, and second, is the response a stereotyped one?

It is clear from a number of studies conducted in Western cultures that the infant cry does not always elicit a response from the caregiver. Bell and Ainsworth (1972) found that primiparous mothers of infants from 0 to 3 months of age ignored a median of 46% of crying episodes. The most responsive mother ignored only 4% of crying episodes and the least responsive ignored 97% of crying episodes. However, lower estimates have been provided by Moss and Robson (Note 2) and Bernal (1972), who found that an average of 17% to 18% of the cries of firstborns were ignored. Mothers of second borns in Bernal's sample ignored fewer crying episodes (8%). The discrepancies between Bell and Ainsworth's estimate and the estimates of the other investigators could be due to the former investigators' having computed a median rather a mean, to different definitions of crying episodes (e.g., Moss and Robson used fusses rather than cries), or to different data collection methods (observational methods vs. diaries).

For those cries that were attended to, the duration of the delay or the latency of the response was reported in two studies. Bell and Ainsworth (1972) reported that mothers delayed a median of 3.83 minutes per hour with a range of 2 minutes to 9 minutes. Bernal (1972) reported that roughly two thirds of the cries were responded to within 10 minutes but that response was delayed for from 10 to 30 minutes for one third of the cries. The smaller range reported by Bell and Ainsworth may be a function of their having computed the duration of the delay per hour rather than per episode.

The Western pattern of often ignoring the cry and delaying response to it contrasts sharply with reports mentioned previously that mothers in hunter–gatherer cultures never ignore infant cries and respond with an average latency of 6 sec (Devore & Konner, 1974). In fact, the findings of Bernal on mothers' unresponsiveness to crying in a Western culture led Richards (1974) to remark that "the important lesson for the infant is how little effect his crying has on his caretakers" (p. 90). On the other hand, several other pieces of evidence would seem to indicate that Richards may have overstated the case. In a small experiment, Moss and Robson (Note 2) wanted to test whether mothers of $3\frac{1}{2}$-month-old infants would respond to a button that lit up (on a schedule prearranged by the experimenters) to indicate that their infants were crying in another room. The investigators had previously observed these mother–infant pairs in the home and had found that mothers responded to 77% of their infants' cries. However, in the button experiment, only 13 out of the 54 mothers responded to the lighted button (as reported by Harper, 1971), a response rate of only 24%.

That the *sound* of the cry may be necessary to convey a sense of urgency was hypothesized by Lenneberg, Rebelsky, and Nichols (1965) in their study of infants born to deaf parents. They reported that deaf parents were not compelled to attend to their crying infants even if they could see the distressed state that their infants were in. It was as though some deaf parents could not tell whether their infants were in distress by looking at them. It appears that the cry may be a necessary though not a sufficient condition for a response in our culture and that a cognitive awareness of distress without the urgency conveyed by the sound itself is ineffective in eliciting a response.

Turning now to whether motor responses to cries can be described as stereotyped, the most frequent intervention in the early months has been found to involve close physical contact (picking up and holding or feeding). Close physical contact accounted for at least half of the interventions in the Bell and Ainsworth (1972) study with primiparas. Of those crying bouts that were attended to, Bernal (1972) observed that 69% resulted in feeding by primiparas and 89% resulted in feeding by multiparas. Furthermore, interventions involving physical contact were 80% effective in soothing the infants in Bell and Ainsworth's sample. These results are consistent with those reported by Korner and Thoman (1970), who found that crying newborns were most effectively soothed and brought to a visually alert state by contact that was combined with vestibular stimulation and an up-

right posture. Over time, however, noncontact interventions such as approaching, vocalization, and other social stimulation became increasingly effective and were more frequently used by mothers (Bell & Ainsworth, 1972; Moss & Robson, Note 2).

Although the studies reported indicate that by far the majority of cries are followed by nurturant responses involving close physical contact, it must not be overlooked that crying is also one of the major precipitants of abuse for children under 12 months of age. In one study of infants battered by their parents, excessive crying was given as the reason for battering by 80% of the parents of infants less than a year old (Weston, 1968).

One could argue, however, that the cry may not always be a sufficient stimulus to elicit any response in our culture, let alone a nurturant response, because of anomalies of development. Lorenz (1965) has stressed that fixed action patterns are very sensitive to "bad rearing" that can cause the disintegration of releasing mechanisms. If the parental attachment system develops optimally under conditions of close parent–infant contact and minimal crying, then a pattern involving little contact and excessive crying may result in nonfunctional or maladaptive responses to the releasing stimulus of crying. In this connection it is worth noting that sensitivity to crying signals typically occurs within a larger cultural pattern of rearing characterized by close mother–infant contact, prolonged breast-feeding, and wide spacing between children. In a survey of 222 cultures, Mead and Newton (1967) found that cultures clustered basically into two types—those with a developed and those with a muted transition between delivery of the infant and the establishment of total physiological separateness. In contrast to the close contact pattern in cultures with the developed transition period, cultures exhibiting the muted transition period, primarily the industrial nations, were characterized by mother–infant separation in the hospital, early weaning, infrequent feedings, a great amount of crying with delayed parental response, and the prevalence of cribs, playpens, and other devices that maintain the baby at a distance and often out of sight of the mother. Because close mother–infant contact, sensitivity to crying, child spacing, and prolonged breast-feeding appear together so frequently in many types of primitive and traditional cultures, Mead and Newton suggested that "there may be strong mechanisms in this interrelation of patterning" (p. 186).

Receptor Mechanisms for the Cry

Lorenz's releaser formulation has inspired researchers to search for the underlying neural mechanisms ("the Holy Grail") involved in stimulus recognition (Brown, 1975). Consistent with a recognition model, researchers of biologically significant sounds have looked for evidence of specialized receptor apparatuses, processing modes, or feature detection capabilities (Wordon & Galambos, 1972). Generally, the search has focused on filtering mechanisms or templates within the auditory processing system. Controversy exists over whether templates consist of one "pontifical" cell or an ensemble of neurons (Wordon & Galambos, 1972) and over whether filters are located centrally or peripherally in the sense organs (Brown, 1975).

Although little neurological research of this nature has been conducted with primates and none with humans, there is a case reported in the literature (Mark & Ervin, 1970) of a teenager with a brain dysfunction who murdered her two sisters when they were babies. When tape recordings of infant crying were played to her, seizurelike activity was recorded in part of her limbic system (amygdala), and the teenager reported a very angry and floating sensation that lasted for several minutes after the cry was turned off. The limbic system is typically considered the "seat of the emotions" (Gellhorn, 1968), but our knowledge of it is limited. It is not clear whether the electrical responses evoked by the cries were determined by general state changes or by the specific acoustic features of the stimulus (Wordon & Galambos, 1972). Relevant to this issue but discussed more fully in a later section is McClean's (1973) suggestion that there may be functional areas in the limbic system that control emotions that guide species-typical behaviors.

Evaluation of the Model of the Cry as a Releaser

From the preceding review of the literature, it is evident that whether one accepts the releaser model of the cry's impact depends on how strictly one interprets the releaser concept. The evidence that most strongly supports the view of the cry as a releaser includes the finding of contagious crying in newborns, the suggestion that deaf parents do not exhibit an urgency about responding to crying, and the near universality of interventions that involve close physical contact. On the other hand, the cry signal does not appear to be a simple and discrete stimulus with a single meaning, but rather its meaning depends upon intensity cues and contextual factors. Cross-cultural variability in responsiveness to crying and the frequent citation of crying as the major factor precipitating abuse suggest that an appeal to additional, though perhaps not incompatible, mechanisms might further increase our understanding of the cry's impact on adults.

Because the releaser formulation originated with observations of invertebrates and lower vertebrates such as insects and frogs, it has been suggested that the IRM *sensu stricto* is not a useful concept when applied to higher vertebrates and particularly to primates (Wilson, 1975). Lorenz (1965), while retaining a fairly strict definition of the concept, has argued for the existence of phylogenetic vestiges of releasing systems in higher organisms (IRMs disintegrated by learning or higher development), but his application of the concept to human behavior in the popular book *On Aggression* (Lorenz, 1966) has been criticized (Piel, 1970).

The approach of a number of comparative psychologists has been to suggest the abandonment of the releaser concept altogether. They point out, for example, that the concept underestimates the role of experience in the ontogeny of fixed action patterns (Lehrman, 1970), that these behaviors are often not as fixed or stereotyped as once thought (Klopfer, 1974), and that the IRM is not a unitary mechanism that corresponds to a discrete center in the central nervous system (Hailman, 1970).

Rather than abandoning the releaser concept, others have dealt with the restrictiveness of the model by altering its definition to extend its meaning. Tinbergen (1948), for example, argued that sign stimuli are not always characterized by simple key features and, furthermore, that releasers need not elicit full-blown motor patterns but only minor elements of behavior or even internal reactions. Also in this tradition, Hinde (1974) suggested that there is a continuum ranging from fixed action patterns that are stable or relatively unaffected by experience to behavior patterns that are labile; because, in this view, most behaviors are affected by both heredity and environment, the question of whether a behavior is *innate* or *acquired* is unanswerable.

The tendency in higher mammals and primates is away from elementary sign stimuli toward signals that do not convey single fixed messages but are graded in intensity and derive their meaning from both intensity cues and contextual factors (Wilson, 1975). For this reason, a model that implicates motivational and cognitive influences in responses to crying is presented in the next section. It should be noted, however, that this model may not be incompatible with the broadened *social releaser* concept used by the English ethologists (e.g., Hinde, 1974; Tinbergen, 1948), who, unlike the ethologists in the classical tradition (cf. Eibl-Eibesfeldt, 1975), tend to view IRMs as motivational entities. Like the releaser model, this cognitive–affective approach can be described broadly as ethological in that it capitalizes upon Bowlby's suggestion (1969) that the cry's impact is releaser-like and that adults' reactions to it may be phylogenetically adapted.

The Cry as an Activator of Emotion

In this section, a model of the cry as an activator of emotions of either an egoistic or an altruistic nature is examined. First, an analogy is drawn between the infant cry and the graded signaling systems of nonhuman primates that function on a motivational as opposed to a symbolic level. Subsequently, an empathy model that implicates both motivational and cognitive factors in responsiveness to the cry is presented.

The Cry as a Graded Signal

In the light of the previous discussions of the physical characteristics of the cry, infant crying can be likened to biologically significant sounds that are graded signals as opposed to discrete signals (Wilson, 1975). Discrete signals operate in an on–off manner with no variation in intensity or duration, whereas graded signals are variable and increase in intensity with the greater motivation of the signaler. In nonhuman primates, discrete signals are found among tree-dwelling primates, and graded signals are common among terrestrial species (Wordon & Galambos, 1972). Discrete signals are adapted to communicating territorial messages over large distances in noisy environments; graded signals are adapted to communication within the troop at relatively close ranges. Graded signals cannot be easily characterized acoustically and are hard to dissect into specific messages. Unlike in human language, these graded signals reflect the motivational state of the signaler rather than state a relationship in symbolic terms (Brown, 1975). The meaning of a graded signal depends on both acoustic cues and nonacoustic cues such as the context in which the signal is employed. However, it is possible that dramatic shifts in the intensity of graded signals may result in shifts in qualitative meaning (Wilson, 1975).

This description of graded signals fits well with the discussion presented earlier concerning the signal value of the cry. It was proposed that cries differ in intensity and that intensity cues may result in listeners' making qualitative distinctions about the underlying causes of cries. In addition, the importance of context in interpreting the meaning of the cry has been emphasized by some researchers (Bernal, 1972; Muller et al., 1974; Wolff, 1969). A cry that occurs immediately after a feed is less likely to be interpreted as a hunger cry than a cry that occurs 2 hours or more after a feed. Thus, both intensity cues and context enter into the interpretation of the cry signal.

There is comparative evidence that supports the conceptualization of the infant cry (as well as some adult sounds such as laughter and moaning) as similar to graded primate signaling systems. These human utterances, which could be described broadly as expletives, are produced with a comparable simplicity and steadiness of upper vocal tract configuration and with predominant variations in the lower vocal tract (Bastian, 1965). According to Bastian, the lower vocal tract controls pitch, timing, and intensity and is closely tied to the autonomic system of arousal. In fact, the anatomical configuration of the infant vocal tract is said to be more similar to that of a nonhuman primate than to that of an adult human (Lieberman, Harris, Wolff, & Russell, 1972). The infant, like the nonhuman primate, does not have a pharyngeal region that can vary in cross-sectional area. Lieberman (1973), noting the similarity in the vocal tracts of infants, adult chimpanzees, and the prehistoric ancestors of *Homo sapiens,* placed great emphasis upon the evolution of the anatomy of the vocal tract as an important factor in the evolutionary development of human language.

That the emission of the infant cry is postulated to be under the control of the hypothalamic and the interrelated limbic system (Chauchard, 1963; Torda, 1976) also reinforces its similarity with nonhuman primate auditory signaling systems. Robinson (1967) has demonstrated that every type of vocalization common to macaques can be elicited by stimulating parts of the limbic system. In fact, removal of areas of the nonhuman primate brain that are homologous to the speech centers in human beings does not affect their vocalization (Myers, 1968). Both the facial and vocal expressions of these monkeys seem not to be under voluntary control, but are primarily controlled by portions of the brain subserving emotional rather than volitional functions. Myer further hypothesized that, based on lesion studies with humans, there is a striking duality in the mode of social communication and its underlying mechanisms in humans. The involuntary expression of affect (facial expressions and vocalizations) appears to be controlled, as in nonhuman primates, by the deep structures constituting the limbic system, whereas the voluntary uses of these expressions are under cortical control.

Thus, in this framework, the cry of the newborn is characterized as an involuntary reflex action to distress that is at first under the control of the hypothalamic/limbic system and only later comes under cortical control. Bell and Ainsworth's (1972) hypothesis that crying is at first reflexive and only later becomes instrumental may reflect increasing cortical control over the emission of the cry in the second half of the first year. Also in this formulation, the cry is regarded as a graded rather than a discrete signal that increases in intensity with the degree of discomfort felt by the infant. Any attempt to interpret the cause of the distress would make use of both acoustic cues of intensity (e.g., dysphonation, hyperphonation, and prolongation of the signal) and contextual factors. This suggests that the meaning of the cry for the listener could be fruitfully studied using a dimensional approach as opposed to the typological approach adopted in the past to study the signal value of cries.

If the infant cry is a graded signal, then the manner of its reception by adults may likewise bear some similarity to the reception of graded signals by nonhuman primates. Effective reception of graded signals is predicated on a modification of the emotional disposition of the listener (Bastian, 1965). Thus, communication with graded signals, both their emission and their reception, can be said to be more on a motivational or emotional basis than on a symbolic one (Brown, 1975). Graded signals do not convey impartial messages but have their effect by influencing the motivational state of the listener.

The emission of the cry as well as its compelling effect on the listener may, then, be under the control of the limbic/hypothalamic system and its modulation of autonomic arousal. In spite of the prominence of cortical control in humans, it is also believed that a considerable role is played by the limbic/hypothalamic system in the reception and execution of elementary types of affective speech and sound making (Chauchard, 1963). Similarly, the triune concept of the brain proposed by McClean (1973) emphasizes the intermeshing of cortical with subcortical functions. The oldest part of the brain comprising the upper brain stem is our inheritance from reptiles and controls functions like respiration and perhaps also stereotyped behavior patterns such as those released by sign stimuli. At the next level, the old mammalian brain includes the hypothalamic/limbic system and controls species-typical behavior such as agonistic, affiliative, and parental behaviors (Altmann, 1966). McClean (1973) presented evidence that the old mammalian brain consists of functional areas that guide behavior with respect to the two life principles—self-preservation and preservation of the species. The old mammalian brain in humans has a similar structure to that found in animals and, he argues, continues to function at an animalistic level in humans with its contribution to the elaboration of emotional feelings that guide behavior. At the highest level, the new mammalian brain or neocortex functions in skilled, discriminatory, and exploratory behaviors.

McClean (1973) hypothesized that the interconnections of the limbic system and the neocortex represent an evolutionary advance in primates that makes possible empathy in terms of both shared affect and a cognitive understanding of another's feelings. The compelling effect of the cry on the listener may be partly due to the fact that it produces an isomorphic response of distress in the listener that is mediated by the limbic system. Before discussing the altruistic basis for responses to the cry, I first examine claims that the motivation to respond to the cry is solely egoistic or self-serving.

The Egoistic Basis for Response to the Cry

In arguments advanced against the releaser model, researchers in the learning tradition (Moss & Robson, Note 2) have suggested that parents respond to the cries of their infants for the same reason that they respond to any noxious sound, that is, to reduce aversive stimulation. This view rests on principles of negative reinforcement as well as on psychophysical assumptions of the relationship between the quality of the auditory experience and the physical characteristics of the sound. Whereas the releaser model emphasizes the uniqueness of a particular stimulus and its receptor mechanisms,

the psychophysical approach stresses general properties of the auditory processing system.

Although no formal psychophysical study has been conducted with the cry sound, Ostwald (1963, 1972, 1973) has presented many speculations as to what physical characteristics of the sound would make it a particularly penetrating noise. He compared the cry to a siren that compresses acoustical energy into a very sensitive region of the auditory spectrum. He reported that the fundamental frequency of most cries is approximately 500 Hz (ranging from 400 to 600 Hz) with heaviest reinforcement at 1000 to 2000 Hz, where the auditory threshold is lowest. In addition, he claimed that the infant cry is one of the loudest sounds human beings ever make, with an average level of 83 to 85 dB at 10 inches (2.54 cm) from the mouth. According to Ostwald, this sound level is 20 dB louder than normal adult speech and is equivalent to the noise of an unmuffled truck. Ostwald's (1963) conceptualization of the basis for parental response to the cry is similar to that advanced by learning theorists:

> One can appreciate why the parent must interfere with the baby's cry: this sound is too annoying to be tolerated beyond a short period of time, particularly at close range. Thus, the cry cries to be turned off! The listener who cannot escape usually reduces the noise by soothing whatever baby needs occasion it. (p. 46)

Psychoacousticians (Kryter, 1970) have claimed, however, that research on the annoyance value of sounds has little application to sounds that convey emotional meanings. The effects of sounds such as the cry that carry information about their sources cannot be quantitatively related to their physical characteristics and are therefore rejected from the concept of perceived noisiness. However, some basic attributes related to perceived noisiness, such as impulsiveness and spectrum content and level, may set fundamental limits on the tolerability of the noise, but emotional meaning can greatly alter tolerance within these limits.

Although the psychophysical approach to the explanation of the perceptual mechanisms underlying the impact of the cry may have some validity, the greatest weakness in this approach is that it accounts best for escape from or avoidance of the crying child and less well for approaches to remove the source of distress. The motivation is egoistic or self-serving in that the parent is motivated to reduce his or her own distress rather than the baby's. Hoffman's (1975) theory of empathic distress as the basis for altruism provides an alternative view of the motivation to respond to the cry.

The Altruistic Basis for Response to the Cry

Many conceptions of empathy have emphasized either the affective aspect of sharing the feelings of others or the cognitive awareness (recognition) of another's plight (Deutsch & Madle, 1975), whereas Hoffman (1975) has presented a formulation that synthesizes the cognitive and affective components in a developmental perspective. At the basis of all altruism is the response of empathic distress or "the involuntary, forceful experiencing of another person's painful emotional state" (p. 613). The experience of empathic distress is necessarily discomforting and unpleasant. As mentioned earlier, on the basis of Simner's (1971) research, Hoffman proposed that empathic distress may exist in a rudimentary form at birth or may come about through classical conditioning.

Integrating the research on cognitive development and on helping behavior, Hoffman proposed three stages in the development of altruistic motivation. Adopting Schacter and Singer's (1962) formulation that the labeling of one's emotion or state of arousal is determined by one's cognitions of the situation, Hoffman suggested that one's cognitive sense of the other likewise determines how one reacts to the distress response of another. Through the development of a cognitive sense of the other, the primitive empathic distress response develops in three stages into a more reciprocal concern for the victim, which Hoffman called sympathetic distress. At first the child is unable to differentiate between his or her own distress and that of another. With the development of the concept of the self as distinct from others, the child's concern for his or her own discomfort is transformed into concern for the other's distress,

but he or she lacks understanding of the cause or remedy of another's distress. At the next stage, the child's attempt to alleviate another's distress is less egocentric and is guided by corrective feedback from immediate situational cues. At the third and highest level, the child can respond not only to situation-specific cues of distress but also to a general representation of the welfare of the victim regardless of the victim's momentary state.

Hoffman further presented some evidence to support the following relationships between altruistic motives and action: (a) Distress cues from another trigger the sympathetic distress response in the observer; (b) the observer's initial tendency is to act; (c) the intensity of the affect and the speed of response should increase with the number of pain cues; (d) if the observer does not act, the observer will continue to experience sympathetic distress or cognitively restructure the situation to justify inaction.

It is instructive to relate Hoffman's formulation to infant crying. First, that the cry is often described as a noxious stimulus accords with the experience of empathic distress as unpleasant. Second, the description of the cry as a compelling stimulus reflects the compulsion to act by observers of another's distress. Third, that the speed of response is increased by the salience of the pain cues has been borne out by Wolff's (1969) data on latency of maternal response to hunger and pain cries. That the intensity of affect should increase with the intensity of pain cues would also be consistent with the concept of the cry as a graded signal.

If, as Hoffman proposed, failure to act results in the observer cognitively restructuring the situation to justify inaction, we may have a clue as to how failures to act and even vengeful acts in response to the cry could come about. At Hoffman's third level of development of altruistic motivation, the observer's representation of the general welfare of the distressed person may override the specific situational cues associated with distress. Thus, if one's child-rearing philosophy is that one should not accede to the infant's distress signals, one might justify inaction under the rubric of teaching the child that he or she must not manipulate the parent. Similarly, conceptions of the relative vulnerability of infants at various ages may lead parents to be less responsive to the distress signals of older infants than to those of younger infants. If, despite this cognitive restructuring, the parents continue to experience involuntary distress, they may try to escape by increasing the distance between themselves and the crying child or, for example, by closing doors to dampen the sound. It may be that continued exposure to the sounds with the attendant involuntary experiencing of a high level of emotional arousal in the parent tips the parent's motivation from altruistic to egoistic, that is, the motivation is no longer to alleviate the infant's distress but to alleviate the parent's distress at having to listen to the sound of crying for prolonged periods of time. This contrasts with the altruistic basis for helping behavior in which the motivation to respond is aroused by another's distress rather than by one's own, in which the major goal is to help the other rather than one's self, and in which gratification is contingent on reducing another's rather than one's own distress (Hoffman, 1975).

The notion that there may be an optimal range of distress cues has been hypothesized by Hoffman (Note 3) and has been referred to as the "critical toxicity" problem by Tompkins (1963) in his discussion of emotions aroused in listeners by infant cries. Distress cues from another must be sufficient to activate distress in the observer but must not be so disturbing as to elicit avoidance of or aggression toward the victim. Excessive and prolonged crying, whether due to constitutional factors in the infant or parental management techniques, may exceed limits of tolerability and overly tax parents' abilities to withstand continuing high levels of emotional arousal.

The model of the cry as an activator of altruistic or egoistic motives in the listener may shed some light on ideologies and actions in our culture with regard to the socialization of crying in infancy. Tompkins (1963) argued on the basis of clinical observation that most people develop articulate philosophies about crying and that, generally speak-

ing, there is a polarization of attitudes such that one is either for or against the crying child. These differing attitudes lead to a polarization of action as well: Either one ignores and thereby punishes the crying child or one tries to soothe the child by removing the source of the distress. In the latter case the parent is motivated to action because he or she experiences sympathetic distress. In the former case, however, the cry is seen as an attempt on the part of the infant to manipulate the parent. The response of the parent is characterized by irritation, anger, and annoyance. By ignoring the cry, these parents hope to reduce the child's dependency and increase his or her self-reliance.

The belief described by Tompkins that babies could be spoiled by responses to their cries was prevalent in child-care advice given to American parents in the first half of this century (Bell & Ainsworth, 1972). To the contrary, Bell and Ainsworth have presented evidence that by not responding promptly to the cries of infants in the first 6 months, parents actually increase the likelihood that their infants will cry more frequently and for longer periods of time in the second half of the first year. By this time, a vicious spiral has been set up whereby more crying leads to more ignoring and more ignoring leads to more crying. In the context of the empathy model proposed, these findings together with Tompkins' observations suggest that one's ideology or cognitive representation of the crying infant's well-being may lead to a caregiving pattern that routinely exposes parents to excessive crying. Frequent exposure to excessive crying would be expected to activate parents' egoistic motives and result in anger toward and avoidance of the source of the sound, thereby completing and escalating the vicious spiral. Reports that abusing parents hold extreme views about spoiling and independence training (Steele & Pollock, 1968) also suggest that their management techniques may foster the excessive crying often given as the reason for abuse.

However, the question can be reversed, and one can ask instead whether prolonged crying may foster the punitive ideology as well as the behavioral unresponsiveness. Bell (1968, 1971) argued that much of the research based on the parent-effect model can be reinterpreted as demonstrating the effects of the child's behavior on the parent. Indeed, recent reports (Lamb, Note 4) suggest that babies who are abused may be especially difficult constitutionally and may therefore precipitate their own abuse. In this regard, it is interesting to consider Bennett's (1971) report that nurses may apply either the altruistic or the punitive ideology described by Tompkins depending on the characteristics of the particular babies in their care. The cries of one very irritable and difficult newborn were viewed as exploitive by nurses, whereas those of an easier baby elicited sympathy and were regarded as legitimate demands. Although both babies were easily soothed by the nurses' interventions, quieting was seen as due to the infant's being "spoiled" in the one case and "socially responsive" in the other. During the first two weeks after delivery, the frequent cries of the spoiled baby were often left unattended, but those of the socially responsive baby were always attended to promptly.

Although extremes of irritability may be the cause of caregiver unresponsiveness in individual cases, the wide variations between cultures in amounts of crying (Konner, 1972; Mead & Newton, 1967) suggest the importance of other factors. Prompt responding to the cry in cultures with a developed transition period in which breast-feeding is the norm may be partly mediated by the effect of the milk letdown reflex on the mother (Mead & Newton, 1967). Greater intimacy of contact (Konner, 1972) in these cultures also provides the opportunity for caregivers to anticipate the needs of babies by interpreting more subtle proximal signals of discomfort. In addition, the ecological setting of many primitive and traditional cultures ensures the availability of many caregivers (Marvin, Van Devender, Iwanaga, LeVine, & LeVine, 1977). Where several caregivers in addition to the mother are concurrently available, prompt attention to distress cues is more feasible than in the typically Western nuclear-family setting with its competing demands on parents. With respect to more cognitive factors, one might hypothesize that in primitive and traditional cultures concep-

tions of the vulnerability of infants are greater than in industrial nations in which infant mortality rates are lower. Cross-cultural variations in responsiveness may be influenced by differing conceptions of the well-being of distressed infants and may result in greater or lesser attention to specific situational cues of distress. That caregivers in a primitive culture interpret the slightest cry as an emergency signal (Devore & Konner, 1974) is consistent with this hypothesis.

The precise interrelationships among empathic distress, parental behavior, motives, ideologies, and infant outcomes require further exploration. In reviewing the available empirical evidence, Hoffman's empathy model has provided useful hypotheses. Future research within this theoretical framework promises to further our understanding of both the compelling effect of the cry and the nature of the outcome for the distressed infant. In addition, it may be useful to supplement this framework with a consideration of the role of nature and nurture in the ontogeny of parental behavior. Given that from an evolutionary perspective parental behavior is not strictly altruistic because it contributes to the parent's reproductive success (Alexander, 1974), one might expect that there are additional mechanisms underlying the display of parental behavior in general and responsiveness to crying in particular. In the next brief section, the role of exposure to infants in the ontogeny of parental behavior is discussed with reference to possible physiological factors that may predispose women to find infantile cues more attractive than men find them.

Factors That Influence the Ontogeny of Parental Behavior

Because care of the young is crucial for genome survival, the question explored in this section is whether there may be mechanisms that foster empathic responsiveness toward infants in particular in addition to the processes outlined by Hoffman (1975) that contribute to a general capacity for empathy. A variety of mechanisms can be accommodated within the framework of Hoffman's model because the theory does not preclude the involvement of constitutional as well as experiential factors in the development of empathy. The interactionist approach to the study of development in behavioral biology also allows for multiple causative factors and has been found useful in organizing the presentation to follow. In this framework, species-typical behavior is seen as developing toward a predictable end product regardless of whether experiential or constitutional factors are involved in its development (Lehrman, 1970; Wilson, 1975).

Although evidence has been cited that women are generally more empathic than men when affective indices are used (Hoffman & Levine, 1976), the question addressed here is whether altruistic behavior toward the young in particular may be facilitated by hormonal action in women. Several findings in the human literature are consistent with the comparative literature in suggesting that hormonal events during the prenatal period and at parturition may enhance the attractiveness of infantile stimulation for women. In humans, there may be further effects of hormonal changes at puberty on the attractiveness of infantile cues for women.

The inductive or irreversible influences of circulating hormones on the sexual differentiation of mammalian fetuses have been documented for the human species as well as for animals (Money & Erhardt, 1972). Regardless of the genetic sex of the fetus, exposure to male sex hormones (androgens) at a critical period results in development according to the masculine anatomical pattern. That these hormonal influences are not merely anatomical is suggested by the research of Money and Erhardt (1972); in a study of fetally androgenized girls, they found that as preadolescents these girls exhibited tomboyish behaviors. When compared with normal girls of their age, the fetally androgenized girls preferred active sports to passive activities such as doll play; they preferred boys to girls as playmates; they preferred to wear functional clothing, for example, slacks rather than dresses; and they were more career- than marriage-oriented in their expectations for their futures. These findings were subsequently replicated with a different sample (Erhardt & Baker, 1974). Money and Er-

hardt speculated that prenatal hormonal exposure to androgens may act to raise the threshold of sensitivity to stimuli associated with the traditional nurturant caregiving role.

Possible influences of hormonal changes at puberty on the attractiveness of infantile cues are suggested by two studies. Fullard and Reiling (1976), in a developmental investigation of Lorenz's "babyness," found that children from about the age of 8 until the age of 12 preferred pictures of adults to pictures of infants. However, between the ages of 12 and 14, the preference of females shifted to infants. Males' preferences for infants did not significantly exceed chance level until adulthood (older than 18 years). Similarly, Huckstedt (1965) found that females preferred a drawing of a supernormal baby (with exaggerated infantile features) to that of a normal baby at 10 to 13 years of age but that this preference was not reliable for males until 18 to 21 years of age.

These findings suggest that prenatal hormonal events and hormonal changes at puberty may operate to bias females toward the eventual adoption of the caregiving role. However, it is difficult to rule out the effects of socialization pressures, as Quadnago and his colleagues (Quadnago, Briscoe, & Quadnago, 1977) have pointed out, or the effects of differential exposure to infants that females may gain prior to adulthood as a result of societal pressures.

Hormonal changes associated with pregnancy and parturition have been held responsible for the induction or rapid onset of caregiving behavior in some mammals, as for example in goats (Klopfer, 1971), in hamsters (Richards, 1966), and in the laboratory rat (Moltz, 1974; Rosenblatt, 1970). These hormonal changes appear to sensitize the mother to cues emanating from the young. Similarly, it has been argued that the hormonal changes surrounding parturition may produce a critical period for the establishment of caregiving in human mothers as well. Salk (1970), for example, found that prolonged separation (1 to 7 days beginning at birth) of mothers from their premature and full-term infants resulted in deviations from the normal pattern of holding infants on the left side (regardless of the handedness of mothers). Klaus and Kennell (1970) also found persistent differences in caregiving behavior between mothers who were allowed early contact with their premature babies and mothers who were only allowed late contact. In more recent studies Klaus has extended his studies to mothers of full-term infants allowed extended or traditional contact with their babies in the first 3 days after delivery. Several findings particularly relevant to responsiveness to infant crying were reported. One month after hospital discharge, mothers in the extended-contact group reported in an interview that they picked up their babies when they cried more often than did mothers in the traditional-contact group (Klaus et al., 1972). Extended-contact mothers were also observed to soothe their infants more often during a physical examination 1 month after hospital discharge (Kennell et al., 1974). The effects of extended contact on responsiveness to crying continued to be evident during a follow-up 11 months later. The mothers in the extended-contact group spent more time soothing their infants when they cried during a physical examination at 1 year of age (Kennell et al., 1974). Klaus and his colleagues argued that early separation affects the mother's commitment and attachment to her infant, her development of a sense of caregiving abilities, and her ability to establish an efficient caregiving regimen (Barnett, Leiderman, Grobstein, & Klaus, 1970). Klaus's findings may have important implications for epidemiological studies of child abuse, as early mother–infant separation along with other adverse factors is reportedly more common among abused children than among unabused siblings (Lynch, 1975).

Whereas hormonal induction in the female rat may be necessary at the outset for the rapid *initiation* of caregiving behavior, *maintenance* of caregiving is less dependent on hormones than on stimulation emanating from the young (Moltz, 1974; Rosenblatt, 1970). The hormonal conditions associated with lactation are not necessary to maintain caregiving behavior, but there is evidence that the presence of prolactin in lactating rats may play a facilitative role by reducing the mother's physiological and behavioral responsiveness to stress (Thoman, Conner, &

Levine, 1970). The influence of prolactin on the behavior of human mothers is not known. It is interesting to note, however, that in one study (Bernal, 1972), breast-feeding mothers responded to crying more quickly than did bottle-feeding mothers and were more likely to respond with feeding. Prompt responding to cries by the breast-feeding mother may be mediated by the effect of the cry on the letdown reflex, as suggested by Mead and Newton (1967), and/or by the effect of the cry on changes in the temperature of the lactating breast (Vuorenkoski, Wasz-Hockert, Kiovisto, & Lind, 1969). Alternatively, if prolactin makes human as well as rodent mothers less prone to stress, the threshold of breast-feeding mothers for the activation of egoistic motives may be higher than that of bottle-feeding mothers.

Turning now to the nonhormonal bases for caregiving behavior, the role of exposure to the young has been emphasized in reviews of the ontogeny of caregiving behavior in mammals (Moltz, 1971; Noirot, 1972; Rosenblatt, 1970). For example, enforced exposure to the young elicits appropriate caregiving such as nest building and retrieving in virgin female and male rats. The length of time required to induce caregiving behavior by exposure to pups (about 6 days), however, far exceeds limits that would be adaptive in the natural context, as pups would die without appropriate care for that period of time. It has been concluded from these studies that caregiving behavior is characteristic of both sexes in the species studied. Caregiving is not dependent on physiological changes for its appearance, but the hormonal changes associated with parturition in the female reduce the duration of exposure to the young that is required to effect a change from attacking or avoidance to caregiving behavior. The effects of the mother's parity on caregiving behavior provide further evidence of the effects of exposure to the young. When the natural sequence of hormonal changes at parturition is disrupted, multiparous rats provide adequate care, whereas 50% of primiparous rats do not (Moltz, 1971). Thus, prior exposure to pup stimuli can compensate for disruptions in mechanisms under hormonal control.

Harper (1971) has reviewed studies that suggest that exposure to the young may sensitize higher as well as lower mammals to care-eliciting cues. A dramatic example of this phenomenon was reported by Harlow and his colleagues (Harlow, Harlow, Dodsworth, & Arling, 1966). Although their isolation-reared rhesus mothers were extremely abusive with their first infants, they performed adequately as mothers with their second infants despite the traumatic nature of their earlier experiences. Salk (1970) also noted the compensatory effects of parity for human mothers. He found that the effect of mother–infant separation on which side of the body the infant was held by the mother was overridden if the mother had previously had a child from which she was not separated in the early postpartum period. That past experience in rearing infants influences the attraction of adult rhesus females to neonates was reported by Sackett (1970). Using a "self-selection circus" that allowed the tested animal to approach other animals of varying ages and sexes, he found that multiparous females spent the most time with neonates, followed by primiparous females who, in turn, spent more time with neonates than did nulliparous females.

The parity of human mothers has also been associated with responsiveness to crying in a study by Bernal (1972). Mothers of second borns were less likely to ignore crying and were more likely to respond promptly than were mothers of firstborns. It is possible that exposure to infants increases the tolerability of the sound of crying for parents and reduces the likelihood of avoidance responses. One could also argue that increased responsiveness to second borns resulted from altered conceptions of the needs of infants in the newborn period. However, a cognitive interpretation of these results seems less likely in the light of Bernal's finding that the behavior of multiparas contradicted their intentions as stated in prenatal interviews. Although 85% of multiparas intended to respond to crying only after 10 minutes, 70% of them actually responded to crying within 10 minutes (for primiparas these percentages were 50% and 62%, respectively). Evidence that contingent responsiveness increases with

parity is, however, equivocal. Another study (Cohen & Beckwith, 1977) found a reduction in contingent responsiveness with parity, which the investigators attributed to the competing demands of the older sibling. A study in which the spacing between infants and their older siblings is controlled may resolve the apparent discrepancy in reported relationships between parity and responsiveness to crying.

In summary, the studies reviewed suggest that a major factor in the ontogeny of mammalian parental behavior is exposure to the young and its enhancement of the attractiveness of the young for adult male and female species members. Females, however, may have somewhat of an advantage over males in terms of hormonal mechanisms that sensitize them to infantile cues. The net effect of the interaction of organismic and experiential factors may account for the greater participation of females in infant care among humans as well as among lower mammals. Nevertheless, among human populations, observed sex differences would be expected to vary across cultures depending on the opportunities provided within each culture for males to be exposed to the eliciting effect of infantile stimulation. The lesser sensitivity to crying in Western than in primitive and traditional cultures may partly be a function of a rearing pattern that not only reduces the intimacy of contact between parents and infants in general but also traditionally prohibits extended contact early in the postpartum period when the mother's sensitivity to care-eliciting cues may be greatest. In addition, increases in the attractiveness of infants for pubertal girls may indicate a heightened susceptibility to observational learning in the prereproductive period for which opportunities are often lacking in Western cultures in which infant care usually takes place in the privacy of the small nuclear-family home. Altruistic behavior toward crying infants must then be viewed in the specific context of ontogenetic processes that sensitize adults to infantile cues and enhance the attractiveness of the young for them.

Summary

This article began with the common observation that the infant cry is a compelling sound that elicits actions of either a nurturant or a nonnurturant (even homicidal) nature from adults. Two models of the mechanisms by which the cry has its powerful impact were examined.

The first model examined was that of the cry as a releaser of parental behavior. In this theoretical framework, the cry is viewed as a distress signal that originally evolved, along with other attachment behaviors, to promote proximity between infants and their caregivers. Close proximity between caregiver and infant functioned to protect the infant from predators in the dangerous environment in which the species evolved. To ensure genome survival, reciprocal mechanisms evolved in caregivers to promote immediate and appropriate responses to the cry signal. In this context, it has been hypothesized that the cry may act as a releaser—a key stimulus that acts figuratively to release a fixed motor response from the receiver. The recognition of the signal and the production of the motor response are said to be under the control of a hypothesized neural filtering system referred to as an innate releasing mechanism. The model of the cry as a releaser was examined in the light of the available literature on the physical characteristics of the infant cry and its effectiveness in eliciting parental behavior. It was found that the key features of the cry stimulus that have been identified relate more to stimulus intensity than to stimulus recognition. In addition, the cry was not found to be invariably effective in eliciting caregiving behavior, particularly in Western cultures. This analysis suggested that the data available, though not entirely compatible with the classical view of IRMs, may be compatible with the broadened definition of releasers as motivational entities adopted by some modern ethologists.

A model of the cry as an activator of emotions was then examined. In this formulation, the cry is likened to the graded signals employed for communication by some nonhuman primates. The cry is viewed as an involuntary reflex action to distress that in-

creases in intensity with the greater motivation or discomfort felt by the infant. The response to the cry likewise consists of an isomorphic response of distress in the observer. Egoistic or self-serving motives for responses to the cry, that is, to reduce the parent's own distress, may account for attempts to avoid or escape from the crying infant. On the other hand, altruistic motives, that is, to reduce the baby's distress, may underlie parental responses aimed at removing the source of the infant's discomfort.

Within the framework of Hoffman's empathy model (1975), altruistic behavior toward persons in distress is viewed as a joint product of the capacity for shared affect and the development of a reciprocal concern for others. Although the triggering of empathic distress usually results in the performance of an altruistic act, exposure to excessive crying may transform the parent's motivation from altruistic to egoistic, that is, to a concern to alleviate the parent's own distress rather than the infant's. The roles of constitutional irritability in the infant as well as parental child-rearing philosophies and management techniques were discussed in relation to the activation of egoistic motives and behavior toward crying infants. Egoistic motives may occur more frequently in Western than in primitive and traditional cultures because of a child-rearing pattern that promotes excessive crying and thereby overly taxes parents' abilities to withstand continuing high levels of emotional arousal.

Because caregiving behavior is important for genome survival, it was argued that the explanatory power of the general empathy model proposed could be further increased by taking into consideration specific ontogenetic processes that foster altruistic behavior toward infants in mammals. A review of the human and comparative literature suggested that the ontogeny of mammalian parental behavior may be under the control of two separate but interacting mechanisms. Whereas hormonal events sensitize female species members to infantile cues, exposure to the young has comparable effects for males and females. Sensitization, whether due to hormonal or to experiential factors, enhances the attractiveness of the young for adult species members. Altruistic behavior toward crying infants must be viewed within the context of these ontogenetic processes that have evolved to ensure genome survival.

Reference Notes

1. Dolhinow, P., & Olson, D. *Infant caretaking patterns of Old World monkeys: Profile of a multiple caretaker species.* Unpublished manuscript, 1975. (Available from P. Dolhinow, Department of Anthropology, University of California, Berkeley, Calif. 94720.)
2. Moss, H., & Robson, K. *The role of protest behavior in the development of mother-infant attachment.* Paper presented at the meeting of the American Psychological Association, San Francisco, August–September, 1968.
3. Hoffman, M. *Empathy, its functioning and development.* Paper presented at the meeting of the Society for Research in Child Development, New Orleans, March 1977.
4. Lamb, M. *Influence of the child on marital quality and family interaction during the prenatal, perinatal, and infancy periods.* Paper presented at the Conference on the Contribution of the Child to Marital Quality and Family Interaction Through the Life Span, Pennsylvania State University, April 1977.

References

Ainsworth, M. D. S. Object relations, dependency, and attachment: A theoretical review of the mother-infant relationship. *Child Development,* 1969, *40,* 969–1025.

Aldrich, C., Sung, C., & Knop, C. The crying of newly born babies: II. The individual phase. *Journal of Pediatrics,* 1945, *27,* 89–96. (a)

Aldrich, C., Sung, C., & Knop, C. The crying of newly born babies: III. The early period at home. *Journal of Pediatrics,* 1945, *27,* 428–435. (b)

Alexander, R. Evolution and behavior. *Research News,* 1974, *25,* 3–18.

Altmann, J. *Organic foundations of animal behavior.* New York: Holt, Rinehart & Winston, 1966.

Barnett, C., Leiderman, P., Grobstein, R., & Klaus, M. Neonatal separation: The maternal side of interactional deprivation. *Pediatrics,* 1970, *45,* 195–205.

Bastian, J. Primate signalling systems and human language. In I. Devore (Ed.), *Primate behavior: Field studies of monkeys and apes.* New York: Holt, Rinehart & Winston, 1965.

Bell, R. Q. A reinterpretation of the direction of effects in studies of socialization. *Psychological Review,* 1968, *75,* 81–95.

Bell, R. Q. Stimulus control of parent or caretaker behavior by offspring. *Developmental Psychology,* 1971, *4,* 63–72.

Bell, S., & Ainsworth, M. Infant crying and maternal responsiveness. *Child Development,* 1972, *43,* 1171–1190.

Bennett, S. Infant-caretaker interactions. *American Academy of Child Psychiatry Journal,* 1971, *10,* 321–335.

Bernal, J. Crying during the first 10 days of life and maternal responses. *Developmental Medicine and Child Neurology,* 1972, *14,* 362–372.

Blurton-Jones, N. Comparative aspects of mother-child contact. In N. Blurton-Jones (Ed.), *Ethological studies of child behaviour.* Cambridge, England: Cambridge University Press, 1972.

Bowlby, J. The nature of the child's tie to his mother. *International Journal of Psycho-Analysis,* 1958, *39,* 350–373.

Bowlby, J. *Attachment and loss: Vol. 1. Attachment.* New York: Basic Books, 1969.

Brazelton, T. Crying in infancy. *Pediatrics,* 1962, *29,* 579–588.

Brown, J. *The evolution of behavior.* New York: Norton, 1975.

Chauchard, P. Emission and reception of sounds at the level of the central nervous system in vertebrates. In R. G. Busnel (Ed.), *Acoustic behavior of animals.* London: Elsevier, 1963.

Cohen, S., & Beckwith, L. Caregiving behavior and early cognitive development as related to ordinal position in preterm infants. *Child Development,* 1977, *48,* 152–157.

Darwin, C. *The expression of the emotions in man and animals.* Chicago: University of Chicago Press, 1965. (Originally published, 1872.)

Deutsch, R., & Madle, R. Empathy: Historic and current conceptualizations, measurement, and a cognitive theoretical perspective. *Human Development,* 1975, *18,* 267–287.

Devore, I., & Konner, M. Infancy in a hunter-gatherer life: An ethological perspective. In N. White (Ed.), *Ethology and psychiatry.* Toronto, Canada: University of Toronto Press, 1974.

Eibl-Eibesfeldt, I. *Ethology: The biology of behavior* (2nd ed.). New York: Holt, Rinehart & Winston, 1975.

Ekman, P., Friesen, W., & Ellsworth, P. *Emotion in the human face.* New York: Pergamon Press, 1972.

Erhardt, A., & Baker, S. Fetal androgens, human central nervous system differentiation, and behavior sex differences. In R. Friedman, R. Richart, & R. Wiele (Eds.), *Sex differences in behavior.* New York: Wiley, 1974.

Fairbanks, G. An acoustical study of the pitch of infant hunger wails. *Child Development,* 1942, *13,* 227–232.

Fullard, W., & Reiling, A. An investigation of Lorenz's "babyness." *Child Development,* 1976, *47,* 1191–1193.

Gellhorn, E. An attempt at a synthesis: Contributions to a theory of emotion. In E. Gellhorn (Ed.), *Biological foundations of emotion.* Glenview, Ill.: Scott, Foresman, 1968.

Hailman, J. Comments on the coding of releasing stimuli. In L. Aronson, E. Tobach, D. Lehrman, & J. Rosenblatt (Eds.), *Development and evolution of behavior.* San Francisco: Freeman, 1970.

Harlow, H., Harlow, M., Dodsworth, R., & Arling, G. Maternal behavior of rhesus monkeys deprived of mothering and peer association in infancy. *Proceedings of the American Philosophical Society,* 1966, *110,* 58–66.

Harper, L. The young as a source of stimuli controlling caretaker behavior. *Developmental Psychology,* 1971, *4,* 73–88.

Hinde, R. *Biological bases of human social behavior.* New York: McGraw-Hill, 1974.

Hoffman, M. Developmental synthesis of affect and cognition and its implications for altruistic motivation. *Developmental Psychology,* 1975, *11,* 607–622.

Hoffman, M., & Levine, L. Early sex differences in empathy. *Developmental Psychology,* 1976, 12, 557–558.

Huckstedt, B. Experimentelle untersuchungen zum "kindchenschema." *Zeitschrift für Experimentelle und Angewandt Psychologie,* 1965, *12,* 421–450.

Illingsworth, R. Crying in infants and children. *British Medical Journal,* 1955, *1*(4905), 75–78.

Irwin, O. Infant speech: Development of vowel sounds. *Journal of Speech and Hearing Disorders,* 1948, *13,* 31–34.

Izard, C. *The face of emotion.* New York: Appleton-Century-Crofts, 1971.

Kennell, J., et al. Maternal behavior one year after early and extended post-partum contact. *Developmental Medicine and Child Neurology,* 1974, *16,* 172–179.

Klaus, M., et al. Maternal attachment: Importance of the first postpartum days. *New England Journal of Medicine,* 1972, *286,* 460.

Klaus, M., & Kennell, J. Mothers separated from their newborn infants. *Pediatric Clinics of North America,* 1970, *17,* 1015–1037.

Klopfer, P. Mother love: What turns it on. *American Scientist,* 1971, *59,* 404–407.

Klopfer, P. *An introduction to animal behavior: Ethology's first century* (2nd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Konner, M. Aspects of a developmental ethology of a foraging people. In N. Blurton-Jones (Ed.), *Ethological studies of child behaviour.* Cambridge, England: Cambridge University Press, 1972.

Korner, A., & Thoman, E. Visual alertness in neonates as evoked by maternal care. *Journal of Experimental Child Psychology,* 1970, *10,* 67–78.

Kryter, K. *The effects of noise on man.* New York: Academic Press, 1970.

Lakin, M. Personality factors in mothers of excessively crying (colicky) infants. *Monographs of the Society for Research in Child Development,* 1957, *22*(1, Serial No. 64).

Lehrman, D. Semantic and conceptual issues in the nature-nurture problem. In L. Aronson, E. Tobach, D. Lehrman, & J. Rosenblatt (Eds.), *Development and evolution of behavior.* San Francisco: Freeman, 1970.

Lennane, K., & Lennane, R. Alleged psychogenic dis-

orders in women: A possible manifestation of sexual prejudice. *New England Journal of Medicine,* 1973, *288,* 288–292.

Lenneberg, E., Rebelsky, F., & Nichols, I. The vocalization of infants born to deaf and hearing parents. *Human Development,* 1965, *8,* 23–37.

Lieberman, P. On the evolution of language: A unified view. *Cognition,* 1973, *2,* 59–94.

Lieberman, P., Harris, K., Wolff, P., & Russell, L. Newborn infant cry and nonhuman primate vocalization. *Journal of Speech and Hearing Research,* 1972, *14,* 718–727.

Lorenz, K. The companion in the bird's world. *Auk,* 1937, *54,* 245–273.

Lorenz, K. *Evolution and modification of behavior.* Chicago: University of Chicago Press, 1965.

Lorenz, K. *On aggression.* New York: Harcourt, Brace & World, 1966.

Lynch, M. Ill-health and child abuse. *Lancet,* August 16, 1975, pp. 317–319.

Lynip, A. The use of magnetic devices in the collection and analysis of the preverbal utterances of an infant. *Genetic Psychology Monographs,* 1951, *44,* 221–262.

Mark, V., & Ervin, F. *Violence and the brain.* New York: Harper & Row, 1970.

Marvin, R., Van Devender, T., Iwanaga, M., LeVine, S., & LeVine, R. Infant-caregiver attachment among the Hausa of Nigeria. In H. M. McGurk (Ed.), *Ecological factors in human development.* Amsterdam: North-Holland, 1977.

McClean, P. *A triune concept of the brain and behavior.* Toronto, Canada: University of Toronto Press, 1973.

Mead, M., & Newton, N. Cultural patterning of perinatal behavior. In S. Richardson & A. Guttmacher (Eds.), *Childbearing: Its social and psychological aspects.* Baltimore, Md.: Williams & Wilkins, 1967.

Milowe, I., & Lourie, R. The child's role in the battered child syndrome. *Society for Pediatric Research,* 1964, *65,* 1079–1081.

Moltz, H. The ontogeny of maternal behavior in some selected mammalian species. In H. Moltz (Ed.), *The ontogeny of vertebrate behavior.* New York: Academic Press, 1971.

Moltz, H. Some mechanisms governing the induction, maintenance, and synchrony of maternal behavior in the laboratory rat. In W. Montagna & W. Sadler (Eds.), *Reproductive behavior.* New York: Plenum Press, 1974.

Money, J., & Erhardt, A. *Man and woman, boy and girl.* Baltimore, Md.: Johns Hopkins University Press, 1972.

Muller, E., Hollien, H., & Murry, T. Perceptual responses to infant crying: Identification of cry types. *Journal of Child Language,* 1974, *1,* 89–95.

Myers, R. Neurology of social communication in primates. *Proceedings of the Second International Congress of Primatology,* 1968, *3,* 1–9.

Noirot, E. The onset of maternal behavior in rats, hamsters, and mice: A selective review. In D. Lehrman, R. Hinde, & E. Shaw (Eds.), *Advances in the study of behavior* (Vol. 4). New York: Academic Press, 1972.

Ostwald, P. *Soundmaking: The acoustic communication of emotion.* Springfield, Ill.: Charles C Thomas, 1963.

Ostwald, P. The sounds of infancy. *Developmental Medicine and Child Neurology,* 1972, *14,* 350–361.

Ostwald, P. *The semiotics of human sounds.* The Hague, Netherlands: Mouton, 1973.

Piel, G. The comparative psychology of T. C. Schneirla. In L. Aronson, E. Tobach, D. Lehrman, & J. Rosenblatt (Eds.), *Development and evolution of behavior.* San Francisco: Freeman, 1970.

Quadnago, D., Briscoe, R., & Quadnago, J. The effect of perinatal gonadal hormones on selected nonsexual behavior patterns: A critical assessment of the nonhuman and human literature. *Psychological Bulletin,* 1977, *84,* 62–80.

Richards, M. Maternal behavior in the golden hamster: Responsiveness to the young in virgin, pregnant, and lactating females. *Animal Behavior,* 1966, *14,* 310–313.

Richards, M. First steps in becoming social. In M. Richards (Ed.), *The integration of a child into a social world.* Cambridge, England: Cambridge University Press, 1974.

Ringel, R., & Kluppel, D. Neonatal crying: A normative study. *Folia Phoniatrica,* 1964, *16,* 1–9.

Robinson, B. Vocalization evoked from the forebrain in Macaca mulatta. *Physiology & Behavior,* 1967, *2,* 345–354.

Rosenblatt, J. Views on the onset and maintenance of maternal behavior in the rat. In L. Aronson, E. Tobach, D. Lehrman, & J. Rosenblatt (Eds.), *Development and evolution of behavior.* San Francisco: Freeman, 1970.

Sackett, G. Unlearned responses, differential rearing experiences, and the development of social attachments by rhesus monkeys. In L. Rosenblum (Ed.), *Primate behavior.* New York: Academic Press, 1970.

Sagi, A., & Hoffman, M. Empathic distress in the newborn. *Developmental Psychology,* 1976, *12,* 175–176.

Salk, L. The critical nature of the post-partum period in the human for the establishment of the mother-infant bond: A controlled study. *Diseases of the Nervous System,* 1970, *31,* 110–116.

Schacter, S., & Singer, J. Cognitive, social, and physiological determinants of emotional state. *Psychological Review,* 1962, *69,* 679–699.

Sedlackova, E. Analyse acoustique de la voix de nouveau-nés. *Folia Phoniatrica,* 1964, *16,* 44–58.

Sherman, M. Differentiation of emotional responses in infants: II. The ability of observers to judge the emotional characteristics of the crying of infants, and of the voice of an adult. *Journal of Comparative Psychology,* 1927, *7,* 335–351.

Simner, M. Newborn's response to the cry of another infant. *Developmental Psychology,* 1971, *5,* 136–150.

Steele, B., & Pollock, C. A psychiatric study of par-

ents who abuse infants and small children. In R. Helfer & C. Kempe (Eds.), *The battered child.* Chicago: University of Chicago Press, 1968.

Stewart, A., et al. Excessive infant crying (colic) in relation to parent behavior. *American Journal of Psychiatry,* 1954, *110,* 687–694.

Stone, L., Smith, H., & Murphy, L. (Eds.). *The competent infant.* New York: Basic Books, 1973.

Thoman, E., Conner, L., & Levine, S. Lactation suppresses adrenal corticosteroid activity and aggressiveness in rats. *Journal of Comparative and Physiological Psychology,* 1970, *70,* 364–369.

Tinbergen, N. Social releasers and the experimental method required for their study. *Wilson Bulletin,* 1948, *60,* 6–51.

Tompkins, S. *Affect, imagery, and consciousness: Vol. 2. Negative affects.* London: Tavistock, 1963.

Torda, C. Why babies cry. *Journal of the American Medical Women's Association,* 1976, *31,* 271–281.

Truby, H., & Lind, J. Cry sounds of the newborn infant. *Acta Paediatrica Scandinavica,* 1965, *163,* 7–59.

Vuorenkoski, V., Lind, J., Wasz-Hockert, O., & Partenen, T. Cry Score: A method for evaluating the degree of abnormality in the pain cry response of the newborn young infant. *Quarterly Progress and Status Report.* Stockholm, Sweden: Speech Transmission Laboratory, Royal Institute of Technology, April 1971.

Vuorenkoski, V., Wasz-Hockert, O., Kiovisto, E., & Lind, J. The effect of the cry stimulus on the temperature of the lactating breast of primipara: A thermographic study. *Experientia,* 1969, *25,* 1286.

Wasz-Hockert, O., Lind, J., Vuorenkoski, V., Partenen, T., & Valanne, E. The infant cry: A spectographic and auditory analysis. *Clinics in Developmental Medicine* (Rep. No. 29). London: Spastics International Medical Publications, 1968.

Wasz-Hockert, O., Partenen, T., Vuorenkoski, V., Michelsson, K., & Valanne, E. The identification of some specific meanings in newborn infant vocalizations. *Experientia,* 1964, *20,* 154.

Wasz-Hockert, O., Partenen, T., Vuorenkoski, V., Valanne, E., & Michelsson, K. Effect of training on ability to identify preverbal vocalizations. *Developmental Medicine and Child Neurology,* 1964, *6,* 397–402.

Weston, J. The pathology of child abuse. In R. Helfer & C. Kempe (Eds.), *The battered child.* Chicago: University of Chicago Press, 1968.

Wilson, E. *Sociobiology: The new synthesis.* Cambridge, Mass.: Belknap Press, 1975.

Winitz, H. Spectrographic investigation of infant vowels. *Journal of Genetic Psychology,* 1960, *96,* 171–181.

Wolff, P. The natural history of crying and other vocalizations in early infancy. In B. Foss (Ed.), *Determinants of infant behavior* (Vol. 4). London: Methuen, 1969.

Wordon, F., & Galambos, R. Auditory processing of biologically significant sounds. *Neurosciences Research Program Bulletin,* 1972, *10,* 1–117.

Received October 27, 1977 ■

CONTENTS (continued)

As of February 1, 1978, Gene V Glass of the University of Colorado succeeded David A. Kenny as Associate Editor for methodological and statistical papers. As soon as most of the methodological articles in an issue have been reviewed by Glass, his name will replace Kenny's on the masthead.

79

Ks
Docket &
issue to the
Library
15/11/79

VOLUME 86 / NUMBER 2 MARCH 1979

Psychological Bulletin

(Continued on inside back cover)

BLISHED BIMONTHLY BY THE AMERICAN PSYCHOLOGICAL ASSOCIATION
ISSN 0033-2909

The *Psychological Bulletin* publishes evaluative reviews and interpretations of substantive and methodological issues in the psychological research literature. The Journal reports original research only when it illustrates some methodological problem or issue. Discussions of methodological issues should be aimed at the solution of some particular research problem on psychology, but should be of sufficient breadth to interest a wide readership among psychologists; articles of a more specialized nature can be directed to the various statistical, psychometric, and methodological journals. The *Bulletin* does not publish original theoretical articles; these should be submitted to the *Psychological Review*.

Abstracts: All articles must be preceded by an abstract of 100–175 words. Detailed instructions for preparation of abstracts appear in the *Publication Manual of the American Psychological Association* (2nd ed.), or they may be obtained from the Editor or from APA Central Office.

Blind review: Because reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the author's identity or affiliation should be on separate pages. Every effort should be made to see that the manuscript itself contains no clues to the author's identity.

Manuscripts: Submit manuscripts in triplicate to the Editor, R. J. Herrnstein, *Psychological Bulletin,* Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, according to instructions provided below.

Instructions to Authors: Authors should follow the directions given in the *Publication Manual of the American Psychological Association* (2nd ed.). Instructions on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors are requested to refer to the "Guidelines for Nonsexist Language in APA Journals" (Publication Manual Change Sheet 2, *American Psychologist,* June 1977, pp. 487–494) before submitting manuscripts to this journal. All manuscripts should be submitted in duplicate and both copies should be clear, readable, and on paper of good quality. Dittoed copies are not acceptable and will not be considered. Authors are cautioned to carefully check the typing of the final copy and to retain a copy of the manuscript to guard against loss in the mail.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1979: $40 domestic, $42 foreign, $7 single issue. APA member rate: $15. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Dept., APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in January) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Printed in the U.S.A. Second-class postage paid at Arlington, Va., and at additional mailing offices.

Vol. 86, No. 2 March 1979

Psychological Bulletin

Statistical Analysis of Dyadic Social Behavior

Helena Chmura Kraemer
Department of Psychiatry and Behavioral Sciences
Stanford University Medical Center

Carol Nagy Jacklin
Stanford University

A method is described for determining the effects of sex (or any other dichotomous characteristic) from the individual correlated behavioral responses observed in dyadic interactional situations. In the illustration used, the dyads are of three types: girl–girl, girl–boy, and boy–boy. Main effects of sex of subject and of sex of partner and interaction effects are estimated and tested, using a generalization of the matched-pair *t*-test approach. Intragroup correlations between paired responses are examined separately. Alternative procedures of analysis are discussed and compared. The present method is extended to interacting groups larger than dyads when the subject characteristic remains dichotomous (e.g., boys and girls in groups of three, four, etc.). In these cases there are three intragroup correlations of interest for responses within the same interacting groups: boys versus boys, boys versus girls, and girls versus girls.

The person with whom one interacts influences what one does. This interdependency of behavior has psychological, as well as statistical, implications. Unfortunately, researchers have either ignored the interdependency or tried to eliminate it, since certain statistical procedures do not allow for it.

To eliminate the dependency, several approaches can be taken. For example, studies with adults often use confederates to provide a standard against which subjects' behavior can be measured. Not only is this strategy difficult to implement in studies of children but it is also not always clear that results obtained using confederates generalize to groups, all of whose members are naive. Alternatively, the dependency of partners' scores can be eliminated by devising scores that apply only to a group or an interacting pair. These scores are dyadic in nature, that is, they can only be attributed to the dyad and not to the individuals within the dyad (e.g., eye-to-eye contact or tug-of-war scores). With scores that can be given to each individual in the dyad separately, one can avoid the issue of dependency by summing individual scores across partners and using only this pair score. In all three cases—using confederates, using scores dyadic in nature, or using pair-average scores—the statistical problems created by dependency are avoided.

Although some questions can be appropriately answered by such scores, there are

This research was partially supported by National Institute of Mental Health Grant HD-9814-02. The authors wish to thank Sandra Lipsitz Bem, Anne Petersen, and John Martin for their careful reading of the manuscript.

Requests for reprints should be sent to Helena C. Kraemer, Department of Psychiatry and Behavioral Sciences, Stanford University Medical Center, Stanford, California 94305.

many important kinds of questions that cannot be answered without individual measures for both members of a dyad (or an interactive group). First, the nature of the mutual dependency of subject and partner behaviors itself varies across groups (Jacklin & Maccoby, 1978), and this fact is of course lost in studies with confererates or with pair or dyad scores. Second, effects of the partner as stimulus may be lost. For example, children have been shown to behave differently in the presence of a boy as opposed to a girl. In the presence of a girl, both boys and girls are more likely to offer toys and behave in other positive ways. However, in the presence of a boy, both boys and girls are more likely to withdraw (see Jacklin & Maccoby, 1978). Use of a score for each pair would only allow overall comparisons of boy–boy, girl–girl and mixed–sex pairs and would mask these sex-of-partner effects.

Thus, a variety of questions can be asked about dyadic effects, the answers to which depend on observations of each member of a dyad (or more generally, on observations of each member of an interactive group). The examples given here are questions about sex differences in children. Isolating the contributions of sex of subject, sex of partner, and the interaction of these two from one another has been found necessary to an understanding of children's social behavior (see Jacklin & Maccoby, 1978). However, the same problem exists for any characteristic of the subject or partner that is being investigated.

Several studies have looked at the nature of interpersonal behavior in young children when some characteristic of the subjects or partners is of interest. Statistical approaches to the dependency problem have varied. Muste and Sharpe (1947), for example, avoided the issue by presenting only descriptive material. A number of researchers have used pair scores (Eckerman, Whatley, & Kutz, 1975; Nadelman & Schiffler, Note 1; Ross & Hay, Note 2). Others seem to have assumed (considering the statistical tests that they used) that partners' behaviors are not interrelated (Doyle, 1975; Hartup, 1974; Langlois, Gottfried, & Seay, 1973; Roff & Roff, 1940; Langlois & Downs, Note 3). The statistical procedures commonly used to evaluate data of this type depend for their validity on adherence to the assumption that the individual scores are independent. However, it is clear that when the behaviors of subjects and their partners are put into the same analysis, this assumption of independence is often violated. As an example, let us consider an extreme case. Suppose that a child's behavior exactly matches that of his or her playmate in some respect. Putting the two members of a pair into an analysis as though they were independent would be equivalent to putting the same case in twice, thus doubling the permissible degrees of freedom and underestimating the error variance. In the more usual case, mutual dependence of behaviors is only partial, but to the extent that mutuality does exist, the inappropriateness of the usual statistical treatments becomes a serious problem. If one wished, for example, to compare the behavior of a girl when she plays with another girl with the behavior of a girl when she plays with a boy, there would be a potentially serious violation of assumptions if both members of each pair were treated as subjects and put into any analysis that disregarded the mutual dependence of behaviors.

This problem is well-known in other contexts. If, for example, one assigns one member of a matched pair to one group and the other to a comparison group, comparison of group response must be based on the matched-pairs t test or the Wilcoxon signed-ranks test rather than on the usual Student's t test or the Mann-Whitney U test, which require independent responses. If one assesses the effect of k different treatments, with each subject exposed to each treatment, analysis of treatment effect must be based on a repeated measures analysis of variance (ANOVA) or a Friedman test rather than one an ANOVA of a one-way layout or a Kruskal-Wallis H test, both of which, again, assume independence of observations. However, in studies of social behavior, the possible dependency has not always been taken into account.

In summary, we suggest that interpersonal behavior often involves the interdependency of subject's and partner's behaviors. Generally, researchers have avoided the statistical problems inherent in this interdependency by using confederates, pair scores, or dyad scores. However, using confederates or dyad or pair

scores has two disadvantages: (a) The nature of the dependency itself, which may be of interest, is often overlooked, and (b) information about partner effects or the interaction of subject and partner effects cannot be obtained.

In the next section, we propose a technique that allows for the analysis of measures of the individual behaviors of both subjects and their partners. The technique involves estimating the level of dependency between subject and partner behaviors and taking this dependency into account before estimating and testing group effects. In general, the analysis parallels the logic of a matched-pairs t test. The approach can be generalized to the examination of individual behavior observed in groups larger than two. We consider (and reject) two alternative approaches that do not take dependency into account and point out the types of errors that this neglect induces, as well as two alternative approaches that may or may not be viable.

Analysis of Individual Behavioral Observations in Dyads

The design considered here is one in which three types of dyads are studied: girl–girl pairs (GG), girl–boy pairs (GB for an observation of the girl in such a pair and BG for an observation of the boy), and boy–boy pairs (BB). The mathematical model used basically involves analysis of variance, fixed effects, and two factors with μ as the grand mean, α as the effect of sex of subject, β as the effect of sex of partner, and γ as the group interaction between sex of subject and sex of partner. Since, however, the pairwise correlations are nonzero and the cell variances may be unequal, the usual analytic procedures may be invalid.

The model states that the paired observations in the girl–girl group are indicated by (x, x') with

$$x = \mu + \alpha + \beta + \gamma + \epsilon,$$
$$x' = \mu + \alpha + \beta + \gamma + \epsilon',$$
$$E(\epsilon) = E(\epsilon') = 0,$$
$$\text{var}\ (\epsilon) = \text{var}\ (\epsilon') = \sigma_{11}^2,$$
$$\text{correlation}\ (\epsilon, \epsilon') = \rho_{11}.$$

Observations of the girl–boy pairs are indicated by (x, y) with

$$x = \mu + \alpha - \beta - \gamma + \epsilon,$$
$$y = \mu - \alpha + \beta - \gamma + \eta,$$
$$E(\epsilon) = E(\eta) = 0,$$
$$\text{var}\ (\epsilon) = \sigma_{12}^2, \quad \text{var}\ (\eta) = \sigma_{21}^2,$$
$$\text{correlation}\ (\epsilon, \eta) = \rho_{12}.$$

Observations of the boy–boy pairs are indicated by (y, y') with

$$y = \mu - \alpha - \beta + \gamma + \eta,$$
$$y' = \mu - \alpha - \beta + \gamma + \eta',$$
$$E(\eta) = E(\eta') = 0,$$
$$\text{var}\ (\eta) = \text{var}\ (\eta') = \sigma_{22}^2,$$
$$\text{correlation}\ (\eta, \eta') = \rho_{22}.$$

Now let $m_1 = (x + x')/2$ in the girl–girl group (the order of assignment within same-sex pairs will be seen to be irrelevant), let $m_2 = (x+y)/2$ and $d_2 = (x-y)/2$ in the girl–boy pairs, and let $m_3 = (y + y')/2$ in the boy–boy pairs. Now m_1, m_2, d_2, and m_3 are random variables with

$$E(m_1) = \mu + \alpha + \beta + \gamma$$

where

$$\text{var}\ (m_1) = \frac{\sigma_{11}^2}{2}(1 + \rho_{11}),$$

$$E(m_2) = \mu - \gamma$$

where

$$\text{var}\ (m_2) = \frac{\sigma_{12}^2 + 2\rho_{12}\sigma_{12}\sigma_{21} + \sigma_{21}^2}{4},$$

$$E(d_2) = \alpha - \beta$$

where

$$\text{var}\ (d_2) = \frac{\sigma_{12}^2 - 2\rho_{12}\sigma_{12}\sigma_{21} + \sigma_{21}^2}{4},$$

$$E(m_3) = \mu - \alpha - \beta + \gamma$$

where

$$\text{var}\ (m_3) = \frac{\sigma_{22}^2}{2}(1 + \rho_{22}).$$

The variances of each of these variables, which are rather complicated functions of the popula-

tion variances and correlation coefficients, can be estimated directly and simply from the sample by the sample variances

$$s_{m_1}^2, s_{m_2}^2, s_{d_2}^2, \text{ and } s_{m_3}^2.$$

Using the above equations, the following are estimators of the parameters:

$$\hat{\mu} = \frac{M_1 + 2M_2 + M_3}{4},$$

$$\hat{\beta} = \frac{M_1 - 2d_2 - M_3}{4},$$

$$\hat{\alpha} = \frac{M_1 + 2d_2 - M_3}{4},$$

and

$$\hat{\gamma} = \frac{M_1 - 2M_2 + M_3}{4},$$

where M_1 is the mean computed over the n_1 averaged responses of the girl–girl pairs, M_2 is the mean computed over the n_2 averaged responses, d_2 is the half difference of the responses of the girl–boy pairs, and M_3 is the mean computed over the n_3 averaged responses of the boy–boy pairs. Since statistics that have different subscripts are independent, the standard errors of $\hat{\mu}$ and $\hat{\gamma}$ are estimated by

$$SE_1 = \frac{1}{4}\left(\frac{s_{m_1}^2}{n_1} + \frac{4s_{m_2}^2}{n_2} + \frac{s_{m_3}^2}{n_3}\right)^{\frac{1}{2}},$$

and the standard errors of $\hat{\alpha}$ and $\hat{\beta}$ are estimated by

$$SE_2 = \frac{1}{4}\left(\frac{s_{m_1}^2}{n_1} + \frac{4s_{d_2}^2}{n_2} + \frac{s_{m_3}^2}{n_3}\right)^{\frac{1}{2}}.$$

If the sample variances are consistent estimators of finite population variances, then for large sample sizes the test statistics $\hat{\mu}/SE_1$, $\hat{\alpha}/SE_2$, $\hat{\beta}/SE_2$, and $\hat{\gamma}/SE_1$ are approximately distributed as standard normal deviates (Cramér, 1946; Lindeberg & Levy Theorem, p. 215; Theorem 20.6, p. 254). To ascertain the statistical significance of any of these effects for large sample sizes, one would compare the magnitude of the corresponding test statistic with the critical values of the standard normal distribution.

To this point, minimal assumptions have been made about the distributions of the data. In particular, we have avoided assuming a bivariate normal distribution or equal variances as well as zero correlation coefficients. It should be noted, however, that if the variances are equal and the correlations are zero, then the statistics above are approximately distributed as t statistics. If the size of degrees of freedom is large, there is little difference in referring to critical values of a t statistic instead of those of the standard normal deviate. For small sample sizes, however, one should be aware of these options as well as of the consequences of inappropriately exercising these options.

If variances are equal and correlations are zero, then the degrees of freedom are $n_1 + n_2 + n_3 - 4$. The procedure in this case is only as valid as the assumptions. If variances are unequal or correlations are nonzero, there is considerable evidence that the nominal significance level of the test may differ substantially from the real significance level (Scheffé, 1959, chap. 10). If the data are nonnormally distributed (but the variance–correlation assumptions are met), it is generally believed that the t test is relatively robust. However, even in this instance, there are special cases that belie this belief (Lee & Gurland, 1977).

Alternatively, one can assume that SE_1^2 and SE_2^2 are linear combinations of independent chi-square statistics, statistically independent of the sample means used to estimate μ, α, β, and γ. In this case, the sample statistics are approximately distributed as t statistics with degrees of freedom that can be estimated from the data (Satterthwaite, 1946). Again, however, if the data are not approximately normally distributed, for small samples the sample variances do not necessarily have chi-square distributions, nor are the sample means and variances necessarily independent. Application of this theory may then be risky.

In addition to these analyses for group effects, intraclass correlation coefficients can be computed separately between paired individuals in the girl–girl and boy–boy dyads, and the product-moment correlation coefficient can be computed for the girl–boy dyad (the intraclass form may be inappropriate if the variances of observations on girl and boy in the GB dyad are not equal). Tests of significance for the two types of correlation coefficients

differ only in specification of degrees of freedom ($n - 1$ for intraclass and $n - 2$ for product moment). Not only can one test for the significance of each correlation coefficient but one can also test for the homogeneity of the three correlation coefficients, and, if they are homogeneous, one can test for the significance of their pooled value (Kraemer, 1975).

A detailed example of the full calculation, estimation, and testing procedures is presented in Tables 1 and 2.

Examination of Alternative Approaches

Inappropriate ANOVA Approach

What are the consequences of applying the usual ANOVA procedure for a balanced design ($n_1, n_3 = n, n_2 = 2n$) and ignoring the paired nature of the data? The parameter estimators are identical to those above and remain unbiased. The variance of these estimators, however, may be either underestimated or overestimated (depending on the true variance–covariance structure) by the use of the mean square error from the ANOVA. For example, if all the variances were equal and all the correlation coefficients were equal to a single nonzero ρ, then the mean square error would estimate σ^2, and the variance of $\hat{\gamma}$ would be taken to be $\sigma^2/8n$. In fact, var $(\hat{\gamma}) = \sigma^2(1 + \rho)/8n$. Thus if $\rho > 0$ (i.e., a positive correlation between subjects and partners), the variance is underestimated, and $\hat{\imath}$ (the interaction between subjects and partners) may appear to be more significant than it is; whereas if $\rho < 0$ (i.e., a negative correlation between subjects and partners), the variance is overestimated, and $\hat{\imath}$ (the interaction between subjects and partners) may appear to be less significant than it truly is. Thus, in general, if the correlation between subject and partner is ignored, incorrect statements may be made regarding the significance of one's findings.

Table 1
Frequency of a Toy Offer in an Interactive Experimental Situation[a]

Raw data			Processed data			
GG	GB, BG	BB	m_1	m_2	d_2	m_3
3, 10	1, 2	2, 1	6.5	1.5	−.5	1.5
5, 0	0, 0	2, 1	2.5	.0	.0	1.5
2, 2	0, 9	1, 1	2.0	4.5	−4.5	1.0
1, 5	3, 2	3, 3	3.0	2.5	.5	3.0
3, 0	5, 2	0, 2	1.5	3.5	1.5	1.0
2, 3	2, 3	2, 0	2.5	2.5	−.5	1.0
2, 3	5, 5	1, 4	2.5	5.0	.0	2.5
3, 0	0, 1	5, 4	1.5	.5	−.5	4.5
2, 7	3, 3	1, 1	4.5	3.0	.0	1.0
2, 3	5, 5	1, 0	2.5	5.0	.0	.5
1, 5	4, 3	1, 4	3.0	3.5	.5	2.5
5, 8	2, 0	0, 0	6.5	1.0	1.0	.0
	1, 2			1.5	−.5	
	3, 14			8.5	−5.5	
	4, 1			2.5	1.5	
	7, 2			4.5	2.5	
	0, 0			.0	.0	
	0, 0			.0	.0	
	0, 0			.0	.0	
	3, 1			2.0	1.0	
	1, 2			1.5	−.5	
M			3.21	2.52	−.19	1.67
SD			1.72	2.16	1.79	1.25

Note. G = girl; B = boy. Intragroup correlation coefficients were .33 (*ns*), .44 ($p < .05$), and −.19 (*ns*) for the GG, GB plus BG, and BB groups, respectively. There was no significant deviation from homogeneity of correlation: $\chi^2 = 3.10$. For the pooled correlation coefficient, $\hat{\rho} = .25$.
[a] Data are from Jacklin and Maccoby (1978).

Table 2
Estimates and Tests of Effects for Frequency of a Toy Offer[a]

Parameter	Estimator	*SE*	*z*	Estimated *df*[b]
μ	2.481	.281	8.829**	34
α	−.480	.249	1.928*	38
β	−.290	.249	1.165	38
γ	−.403	.281	.150	34

[a] Data are from Jacklin and Maccoby (1978).
[b] Using the method described in Satterthwaite (1946)
* $p < .10$.
** $p < .001$; the effect of the mean estimator's being significantly different from zero is trivial in this case, since only positive scores were used. However, if negative and positive scores had been used, the estimator significance might be useful.

Multiple t-Test Approach

One might also consider comparing (e.g., using *t* tests) the responses of boys in same-sex dyads (averaging paired responses) with those

of boys in mixed-sex dyads (BB vs. GB) and similarly consider comparing BB versus BG, BB versus GG, BG versus GG, and GB versus GG). Since subjects' responses in different groups are independent, these would be valid tests. There are five such tests, with the proposed multiple t tests nonindependent. In such a case, the probability that one or another of the tests will exceed the stated significance level by chance may be quite high. However, this is not the crucial point. The main problem lies in the fact that one can extract from these test results no clear evidence as to which of the factors of interest (sex of subject, sex of partner, or interaction) is operative, since each procedure tests some combination of factors. For example, the test comparing the boys' responses in the BB and BG dyads tests whether $\alpha = \gamma$. If the test turns out to be significant, it is not clear whether $\alpha = 0$ and $\gamma \neq 0$, whether $\alpha \neq 0$ and $\gamma = 0$, or whether both are nonzero but equal to different values. Whether or not the sex of partner effect equals the interaction effect is in general simply not of research interest.

Discarding-Data Approach

One may also discard one randomly selected subject from each dyad to ensure independence of the data used in analysis. However, computational simplicity is in this case achieved at considerable loss of power and, in addition, with a loss of information about the dyadic interaction itself, that is, about the pairwise correlations, which may be of interest.

Multivariate Analysis of Variance

Another valid approach to the analysis entails regarding the original sample as one from a bivariate distribution with means and covariance structure as specified. If the distribution is specifically bivariate normal with identical covariance structure in the three groups (BB, BG, and GG), one can use a multivariate ANOVA approach (Anderson, 1958, p. 215–221). For small sample sizes, this procedure is preferable if the assumptions are satisfied, but is risky otherwise.

Generalization to Triads and Beyond

The approach detailed for dyads can be generalized to the situation in which subjects are of two types (girls and boys) but are combined in groups of size three or more. The principles used in the case of dyads are readily extended to larger groups, but the notation becomes cumbersome. We thus sketch only briefly how this generalization can be effected. Table 3 presents a mathematical model for this situation.

Each group comprises k members, and there are $k + 1$ such groups indexed by the number of girls in the group: $j = 0, 1, 2, \ldots, k$. In the model, α is the sex of subject effect, β_j is the main effect of that subject's partnership group (with $k - 1$ members), comprising $j - 1$ girls and $k - j$ boys, and γ_j is the interaction effect between sex of subject and such a partnership group ($\sum \beta_j = \sum \gamma_j = 0$; $j = 1, 2, \ldots, k$). Thus one notes that the response of a boy in a group of $j - 1$ girls and $k - j + 1$ boys is parameterized in terms of the same parameters (β and γ) as that of a girl in a group of j girls and $k - j$ boys.

We assume that the expected response of all girls in any group is the same, as is that of all boys in a group. Thus each k-dimensional response vector, for purposes of estimating the

Table 3
Model for Individual Responses in Groups of Size k

Index j	Type of group: No. Girls	Type of group: No. Boys	Individual response: Girls	Individual response: Boys	No. Groups Observed	Mean response: Girls	Mean response: Boys
0	0	k	—	$\mu - \alpha + \beta_1 - \gamma_1 + \epsilon'$	n_0	—	$\bar{Y}_{\cdot 1}$
1	1	$k - 1$	$\mu + \alpha + \beta_1 + \gamma_1 + \epsilon$	$\mu - \alpha + \beta_2 - \gamma_2 + \epsilon'$	n_1	$\bar{X}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	k	0	$\mu + \alpha + \beta_k + \gamma_k + \epsilon$	—	n_k	$\bar{X}_{\cdot k}$	—

parameters, can be reduced to a bivariate vector $(X_{ij}, Y_{i(j+1)})$ in which X_{ij} is the average response of girls in the ith group of type j $(i = 1, 2, \ldots, n_j)$ and $Y_{i(j+1)}$ is the average response of boys in such a group. Thus no matter what the size of the group, the problem can be reduced to that of considering a bivariate response.

Within a group there are only three intersubject correlation coefficients of interest: that of the error components between a pair of girls, that between a pair of boys, and that between a boy and a girl. When these correlation coefficients are nonzero, as one would expect, X_{ij} and $Y_{i(j+1)}$ are correlated responses.

If the data are reorganized in terms of a two-way layout (Sex of Subject × Type of Partnership Group), the process of obtaining unbiased estimators of the parameters is clear (see Table 4).

Table 4
Means Computed From Groups of Size k

Sex of subject	No. of girls in partnership group: 0	1	2	⋯	⋯ $k-1$	Mean
Girl	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{.3}$	⋯	⋯ $\bar{X}_{.k}$	$\bar{X}_{..}$
Boy	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{.3}$	⋯	⋯ $\bar{Y}_{.k}$	$\bar{Y}_{..}$
Mean	$\bar{Z}_1$	$\bar{Z}_2$	$\bar{Z}_3$		$\bar{Z}_k$	M

Note. Each mean in the first row is correlated with the mean in the second row that is one column to the right (e.g., $\bar{X}_{.1}$ with $\bar{Y}_{.2}$); $\bar{X}_{.j} = \sum_i X_{ij}/n_j$, $\bar{Y}_{.j} = \sum_i Y_{ij}/n_{j-1}$, $\bar{X}_{..} = \sum_j \bar{X}_{.j}/k$, $\bar{Y}_{..} = \sum_j \bar{Y}_{.j}/k$, $M = (\bar{X}_{..} + \bar{Y}_{..})/2$, and $\bar{Z}_j = (\bar{X}_{.j} + \bar{Y}_{.j})/2$.

Thus

$$\hat{\mu} = M,$$

$$\hat{\alpha} = \bar{X}_{..} - M,$$

$$\hat{\beta}_j = \bar{Z}_j - M \quad \text{where} \quad j = 1, 2, \ldots, k, \text{ and}$$

$$\hat{\gamma}_j = \bar{X}_{.j} - \bar{Z}_j - \bar{X}_{..} + M \quad \text{where} \quad j = 1, 2, \ldots, k.$$

The problem lies in the estimation of standard errors for these estimators, since they comprise both dependent and independent sample means. The procedure is illustrated by the calculation for, say, $\hat{\beta}_2$.

First $\hat{\beta}_2$ is re-expressed in terms of sample means as

$$\hat{\beta}_2 = (\bar{X}_{.2} + \bar{Y}_{.2})/2 - (\sum_{j=1}^{k} \bar{X}_{.j} + \sum_{j=1}^{k} \bar{Y}_{.j})/2k.$$

Correlated means are isolated, that is,

$$\hat{\beta}_2 = (2k)^{-1}\{[(k-1)\bar{X}_{.2} - \bar{Y}_{.3}] + [(k-1)\bar{Y}_{.2} - \bar{X}_{.1}] - \sum_{j=3}^{k-1} (\bar{X}_{.j} + \bar{Y}_{.(j+1)}) - (\bar{Y}_{.1}) - (\bar{X}_{.k})\}.$$

In this case, each bracketed term is independent of any other bracketed term, since the terms arise from different groups. Thus the variance of $\hat{\beta}_2$ is the sum of the variances of the bracketed terms divided by $4k^2$. The variance of any one of these terms, for example, that of

$$[(k-1)\bar{X}_{.2} - \bar{Y}_{.3})],$$

is obtained as follows.

For each of the n_2 groups observed comprising two girls and $k - 2$ boys we compute

$$U_i = (k-1)X_{i2} - Y_{i3}, \quad \text{where} \quad i = 1, 2, \ldots, n_2.$$

The variance estimate of σ_U^2 is then the sample variance of these observations, s_U^2. The term of interest,

$$(k-1)\bar{X}_{.2} - \bar{Y}_{.3} = \bar{U},$$

has population variance σ_U^2/n_2, estimated by s_U^2/n_2.

In this way the variance estimates of each of the bracketed terms are compiled and combined to yield the variance estimate of $\hat{\beta}_2$. Once again, if the sample variances are consistent estimators of finite population variances, the large sample distribution of the ratio of the statistic to its standard error under the null hypothesis is approximately standard normal. As above, under certain circumstances the small sample distribution of these statistics may be either exactly or approximately a t distribution.

Summary

The study of social behavior without confederates is complicated by the dependency of subjects' and partners' behaviors. We have suggested a statistical approach that takes

into account this dependency. Although we have illustrated the approach with data from dyads, the method can be generalized to groups of any size.

We have discussed the problems encountered with alternative statistical analyses. The most common approach in the literature is to use an ANOVA without taking into account the dependency of the data. Even if all the other ANOVA assumptions are met (equal variances and balanced design), the test for the interaction term, for example, will be biased depending on the amount and direction of dependency (correlation) that exist. Positive correlations lead to an overestimate of the significance of the ANOVA interaction term; negative correlations lead to an underestimate of the significance of the ANOVA interaction term. A second approach we discussed was the multiple *t*-test approach. The problem with this approach is that it does not give clear evidence of which factors of interest account for the results. A third approach is to drop half the subjects from the analysis to eliminate the dependency. One problem with this approach is a loss of power of the test used, but perhaps a more serious problem is that the actual correlations between subject and partner may be lost.

Finally, we note that all of these approaches represent attempts to reduce what is in general a complex multivariate analysis problem to a series of relatively simple univariate analysis problems. In certain circumstances (multivariate normality and certain covariance structures) the multivariate approach is not so complex as to discourage its use and in these cases is preferable.

Interdependence of subjects' and partners' behaviors reflects the real social situation. In some cases, if we are to understand the dynamics of social behavior, the use of confederates may not be the best choice, even when it is possible. The analytic procedures, however, should be appropriate to the nature of the data.

Reference Notes

1. Nadelman, L., & Shiffler, N. *The influence of sex of dyads on children's cooperation-competition in the scissors-paper game.* Paper presented at the meeting of the Midwestern Psychological Association, Detroit, May 1977.
2. Ross, H., & Hay, D. *Conflict and conflict resolution between 21-month-old peers.* Paper presented at the meeting of the Society for Research in Child Development' New Orleans, March 1977.
3. Langlois, J. H., & Downs, A. C. *Peer relations as a function of physical attractiveness: The eye of the beholder or behavioral reality?* Paper presented at the meeting of the Society for Research in Child Development, New Orleans, March 1977.

References

Anderson, T. W. *Introduction to multivariate statistical analysis.* New York: Wiley, 1958.

Cramér, H. *Mathematical methods of statistics.* Princeton, N.J.: Princeton University Press, 1946.

Doyle, A. B. Infant development in day care. *Developmental Psychology*, 1975, *11*, 655–656.

Eckerman, C. O., Whatley, J. L., & Kutz, S. L. Growth of social play with peers during the second year of life. *Developmental Psychology*, 1975, *11*, 42–49.

Hartup, W. W. Aggression in childhood: Developmental perspectives. *American Psychologist*, 1974, *29*, 336–341.

Jacklin, C. N., & Maccoby, E. E. Social behavior at 33 months in same-sex and mixed-sex dyads. *Child Development*, 1978, *49*, 557–569.

Kraemer, H. C. On estimation and hypothesis testing problems for correlation coefficients. *Psychometrika*, 1975, *40*, 473–485.

Langlois, J. H., Gottfried, N. W., & Seay, B. The influence of sex of peer on the social behavior of preschool children. *Developmental Psychology*, 1973, *8*, 93–98.

Lee, A. F., & Gurland, J. One-sample t-test when sampling from a mixture of normal distributions. *Annals of Statistics*, 1977, *5*, 803–807.

Muste, M. H., & Sharpe, D. F. Some influential factors in the determination of aggressive behavior in preschool children. *Child Development*, 1947, *18*, 11–28.

Roff, M., & Roff, L. An analysis of the variance of conflict behavior in preschool children. *Child Development*, 1940, *11*, 43–60.

Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 1946, *2*, 110–114.

Scheffé, H. *The analysis of variance.* New York: Wiley, 1959.

Received November 22, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 225–241

Obsessive–Compulsive Personality: A Review

Jerrold M. Pollak
Boston College

A review of the term *obsessive–compulsive personality* (or *anal character*) is presented. Statistical studies of obsessive–compulsive personality conducted over approximately the last two decades are then reviewed, emphasizing the extent to which they support theory, clinical observation, and description. Evidence is still needed on precise etiological determinants, and persuasive evidence in favor of classical psychoanalytic theories about etiology is lacking. Empirically based findings to date, however, are congruent with clinical observation, description, and prediction regarding the salient behavioral characteristics and character styles of obsessive–compulsive individuals.

The obsessive–compulsive personality or *anal character*, as it is sometimes termed, came under close scrutiny by Freud and his colleagues (Abraham, 1921/1953; Freud, 1908/1963; Jones, 1918/1961). Carr (1974) credited Esquirol (cited in Carr, 1974), who worked in the early part of the 19th century, with the first publication relating to compulsions and similar phenomena. According to the early Freudians, the anal character, as they referred to it, arises out of conflicts between parents and child over bowel training in the 2nd to 3rd year of life. The introduction of bowel training was thought to bring with it an inevitable conflict between the child's desire to freely manipulate elimination (expulsiveness) and retention (retentiveness) and the need of the child's primary caretakers to regulate their child's anal activities and expressions in line with prevailing cultural and societal standards of cleanliness and impulse control. If the primary caretakers prove to be too punitive, impatient, and intolerant of their charge's willfulness and autonomy, or if the training comes either too early or too late or is experienced as inordinately frustrating or gratifying, the inevitable struggle over bowel control intensifies. According to classical analytic theory, this may very well establish the groundwork for anal fixations and hence the development of a predominantly anal or obsessive–compulsive character structure in the child.

For Freud and for all later analytic and nonanalytic clinicians, obsessional personality or anal character is to be carefully distinguished from the psychiatric syndrome of what has been called obsessive–compulsive or obsessional neurosis. In an obsessional neurosis the individual suffers from persistent intrusion of undesired thoughts (obsessions) or urges and actions (compulsions) that he or she finds extremely difficult, if not impossible, to control and ultimately stop. Following Salzman (1968), the obsession may be viewed as a persistent ritualized thought pattern, whereas the compulsion is a persistent ritualized behavior pattern. Either or both may be salient in the clinical picture, and anxiety and distress usually are concomitants of the disorder.

Individuals with a predominantly obsessive–compulsive personality are considered asymptomatic, that is, what defines the individual consists of particular constellations of traits, defenses, and life-style and not the presence of psychiatric symptomatology.

While obsessive–compulsive neurosis is a relatively rare phenomenon (Templer, 1972), obsessive–compulsive personality or anal char-

Appreciation is extended to Veronika Neudachin for her editorial assistance.

Requests for reprints should be sent to Jerrold M. Pollak, Department of Counseling Psychology, McGuinn Hall, 3rd Floor, Boston College, Chestnut Hill, Massachusetts 02167.

acter structure is not. In fact, one could argue that in Western culture obsessive–compulsive personality is one of the, if not the predominant, social character structures, embodying as it does so much of the general world view of the Protestant Work Ethic and capitalist social and economic organization (Honigmann, 1967).

Obsessive–compulsive personality has also been observed to prevail in some non-Western but relatively advanced industrial societies, such as Japan (Gorer, 1943), and is very much in evidence within various professional groups and specialty occupations and among individuals in bureaucratic and managerial positions. It could be said as well that many of the traits characteristic of obsessional personalities, for example, perseverance, industriousness, thriftiness, ambition, self-control, and so on, are highly regarded and rewarded within capitalistic, technological societies, serve to promote in their possessors feelings of self-worth and acceptability, and generally provide them with a foundation for emotional stability and relative resistance to stress (Paykel & Prusoff, 1973).

Freud (1908/1963) delineated one particular constellation of traits, namely, obstinacy, parsimony, and orderliness, that constitute the core of what he termed the *anal retentive* or *anal character* type and that arise from sublimations of and reaction formations against infantile anal erotic impulses that press for expression. In Freud, orderliness refers to both exceptional bodily cleanliness and a high degree of reliability and conscientiousness in the performance of all actions, however inconsequential. Parsimony involves frugality and, in the extreme, stinginess and avarice, whereas obstinacy involves strong tendencies to be negativistic, defiant, and even hostile in relation to authority figures. According to Freud, orderliness develops from the internalization of parental demands for bowel control, whereas parsimony develops from the continuation of the infantile tendency to retain feces, both because of the erotic pleasure that accompanies retention and because of the fear of losing the overvalued product. Freud viewed obstinacy as developing from resistance to parental demands. In the traditional psychoanalytic model, considerable aggression, that is, anal sadistic impulses and attitudes, is generated in the child as a result of his or her struggle to insure autonomy. When this struggle leads to fixations or developmental arrests at the anal stage, the infantile unresolved aggression is expressed in later life in any number of ways, for example, in passive–aggressive withholding behavior (or other indirect means of defying the wishes or dictates of others) or in the adoption of extremely conventional, oversocialized or reactionary attitudes via reaction formation and identification with the aggressor. The adoption of reactionary attitudes is oftentimes linked with an aggressive, hypercritical, and controlling attitude toward others. Regardless of the exact form taken by infantile aggression in specific individuals, ambivalent attitudes toward the expression of hostile feelings and impulses particularly, but also toward all manifestations of the affect and impulse life, are thought to be paramount in the personality makeup of the adult obsessive.

The traits, defenses (undoing, reaction formation, intellectualization, and isolation), and life-styles viewed as characteristic of anal characters or obsessive–compulsive personalities have been further elaborated by later analysts with more refined theoretical formulations, case history data, and analyses of responses to projective tests like the Rorschach test (Fenichel, 1945; Rado, 1959; Schafer, 1954; Shapiro, 1965). Despite some ambiguities and inconsistencies in the voluminous clinical literature, there appears to be considerable consistency in the personality descriptions that emerge. This is true even when one compares psychoanalytic descriptions of the anal character with descriptions of the obsessive–compulsive character of obsessive personality that emerge from less heavily psychoanalytically influenced theoreticians and practitioners. The latter accept the analytic description of the character, but do not necessarily agree with its etiologic assumptions concerning psychogenesis in the anal stage of psychosexual development. Ingram (1961a), for instance, compared descriptions of the obsessive personality found in leading psychiatric texts with descriptions of the anal character given in the classical psychoanalytic

papers of Freud, Abraham, and Jones and found highly similar descriptions with numerous features in common, for example, the characteristics of orderliness, persistence, and rigidity. For descriptive purposes Ingram felt there was no point in distinguishing between the two terms. The terms *obsessional*, *obsessive*, or *obsessive–compulsive* personality appear to be the preferred terms insofar as no etiologic assumptions are implied by them.

In his review of descriptive features, Ingram emphasized the following composite description of obsessive individuals:

People who manifest *obstinacy* are persevering, thorough, reliable, overconscientious, and have considerable qualities of endurance and drive. They are dogged and persistent. All of this may lead them to be called obstinate, stubborn, or defiant. Added to this is a quality of self-willed independence that leads to a desire to do things their own way and to a dislike of interference. They are rigid and inflexible, particularly in terms of ideals, values, and standards of conduct for themselves as well as others. They are not emotionally demonstrative and may be cold, aloof, and distant. They are often critical, controlling, and in the extreme, power loving.

Individuals who demonstrate *inconclusiveness* are indecisive, vacillate in behavior and thought, cannot leave well enough alone, and make unsatisfying attempts to reach order and perfection. They are always busy and never finished and feel harassed by responsibilities and obligations. They repeat and check their work repeatedly and needlessly, and they continuously weigh the pros and cons of decisions. They fear error, are afraid of making mistakes or omissions, have strict moral scruples, and fear violating the social code. In general, they are uncertain of themselves, are insecure, are prone to worry and doubt, and can be hesitant; but they often try to conceal these traits.

People who display *orderliness* are overorderly, live by routine, and become easily upset when their routine is disturbed by unforeseen events or circumstances. They are meticulous, perfectionistic, and sticklers for precision. They are fond of indexing, tabulating, organizing, and planning and crave accuracy, symmetry, and rationality.

Individuals who exhibit *parsimony* are frequently avaricious, particularly with regard to possessions, money, and time, all of which are not to be wasted.

The review of the literature presented in the following pages focuses on statistically based research studies of obsessive–compulsive personality conducted over the past two decades or so that are judged to be relevant to the descriptive, psychogenetic, and psychodynamic considerations that have emerged primarily from theory and clinical observation.

Empirical Studies Relevant to the Etiology of Obsessive–Compulsive Personality

A number of investigators have sought to substantiate relationships between toilet training practices and the development of anal character traits as originally proposed by Freud (e.g., Beloff, 1957; Bernstein, 1955; Durrett, 1959; Finney, 1963; Hetherington & Brackbill, 1963; Holway, 1949; Huschka, 1942; Kline, 1969; D. R. Miller & Swanson, 1966; Sears, Rau, & Alpert, 1965; Sewell, Mussen, & Harris, 1955; Straus, 1957; Whiting & Child, 1953). The majority of these studies have focused largely on the age toilet training was initiated, the age it was brought to completion, and the degree to which it may have been inordinately lax or severe. The designs typically involved the collection of retrospective accounts, usually by mothers, of the toilet training period in an attempt to relate these parental recollections of when and how training proceeded to the degree of the offspring's anal orientation, as measured by teacher and parent ratings, response to anality questionnaires, and performance on various behavioral tests (S. Fisher & Greenberg, 1977).

A review of these studies offers, at best, meager support for the hypothesized relationship between toilet training practices and the development of anal or obsessive–compulsive character structure. Orlansky (1949) concluded that knowledge of sphincter training was insufficient to substantiate or disprove the Freudian position. Sewell et al. (1955) attempted to relate toilet training practices to personality assessments made at

5 years of age, but found no evidence for the Freudian view. Beloff (1957) found that the personality traits of orderliness, parsimony, and obstinacy occurred together, but could not find significant relationships between how coercively toilet training was carried out, as assessed through interviews of mothers of college students, and student and peer ratings of anal characteristics. In a review of child-rearing practices, O'Connor and Franks (1960) reported no conclusive evidence for the Freudian interpretation. Studies subsequent to that review (e.g., Hetherington & Brackbill, 1963; Finney, 1963; Sears et al., 1965) also indicated that reported parental accounts of toilet training practices were not related to ratings of children's anal characteristics. Moreover, in a series of studies, Gottheil and Stone (1968, 1974) and Stone and Gottheil (1975) could find, at best, only a "slight preferential association" between anal personality patterns and patterns of bowel habits in samples of normal, neurotic, and psychosomatically involved outpatients, that is, bowel habit questionnaire items did not load highly on the anal trait factor that was identified. Kline (1968), however, did report significant positive relationships between a measure of anal eroticism derived from the Blacky Pictures (Blum, 1949) and several measures of obsessional traits and symptoms (viz., measures from Sandler & Hazari, 1960, and Beloff, 1957, and his own AI3 Scale of Anality).

Although it appears that there is little, if any, empirical evidence for the classical psychoanalytic position on the etiology of the obsessive–compulsive or anal character type, there are suggestions in clinical observation and in some of the same studies that failed to support the Freudian point of view of relationships between anal character traits in the child and the existence of anal characteristics in the parents (Beloff, 1957; Finney, 1963; Hetherington & Brackbill, 1963). Beloff administered a questionnaire concerned with anal traits to a sample of male and female college students and their mothers and found positive relationships between parental anal orientation and the anal orientation of the children.

Hetherington and Brackbill administered a questionnaire measuring Freud's anal triad, namely, obstinacy, orderliness, and parsimony, to a sample of fathers and mothers. Their male and female children's anality was assessed by performance on a series of tasks measuring anal behaviors such as parsimony, obstinacy, and perseverance. Significant positive correlations were found between degree of anality in mother and daughter but not in mother and son. No significant relationships were found between a father's anality and that of his children of either sex. Speculating that strict toilet training is simply one expression of a more general pattern of parental rigidity, Finney found, as predicted, that clinicians' ratings of general rigidity in a sample of mothers bore a significant positive relationship to the degree of the child's anal orientation.

The results of these studies, which are indicative of comparable degrees of anal orientation in children and their parents, suggest a number of possibilities regarding etiology. It may be that obsessive–compulsive personality or anal character structure emerges from repeated contact, association, and clashes throughout childhood between the child and figures such as parents, teachers, and relatives directly involved in caretaking responsibilities, who are themselves fairly rigid, controlling, and generally obsessional in their style of relating to the children in their charge. As Carr (1974) has pointed out, even if a relationship between rigid toilet training practices and obsessional traits could be shown, this association could easily be interpreted as a function of childhood training in general rather than of specific repressive toilet training. This point of view is not inconsistent with the idea that the effect of a rigid, obsessional parental orientation could very well be maximal before or during the toilet training phase, when unresolved anal conflicts in one or both parents are stirred up anew, leading to increased anxiety and more pronounced recourse to obsessional behavior as a defense against the impact of the stressful circumstances (S. Fisher & Greenberg, 1977). This could occur even if the primary caretakers did not begin toilet training particularly early or late. It may be, then, that toilet training practices are not causal in any

strict sense, but are a correlate of a larger and more influential child-rearing pattern. In this view, obsessive–compulsive style is seen as largely socially learned behavior that results from the imitation and modeling of significant others over a number of years throughout the childhood period.

One cannot discount, as well, the possible role played by some as yet vaguely defined constitutional factors in the etiology of obsessive–compulsive personality. Freud (1908/1963) himself made reference to constitutional influences that might result in an especially intense inborn erotic sensitivity in the anal zone that by itself or in interaction with experiences during the toilet training phase results in the development of a predominantly anal adult personality orientation.

In a review of research findings on obsessive–compulsive neurosis, Templer (1972) concluded that, although there is a high incidence of various types of psychopathology among relatives of obsessive–compulsive neurotics, the possible etiological role played by genetic or other constitutional factors is unclear. This point of view is shared by Carr (1974). In the case of the development of obsessive–compulsive personality, the possible role played by constitutional influences still remains in the realm of speculation because, to date, there has been little empirical research in this area. A study by Hays (1972), however, of the family pedigrees of 17 psychiatric patients, mostly female, that carried a diagnosis of psychotic depression and had premorbid obsessive–compulsive personalities did find evidence to support an interaction effect of genetic predisposition, sex of the child, and child-rearing style in the genesis of obsessive–compulsive personality.

Some theorists and clinicians stress the need for obsessive–compulsive personality style as a character defense or armor for the ego against the ambiguities, uncertainties, and anxieties inherent in human existence (e.g., Becker, 1974; M. H. Miller & Chotlos, 1960; Salzman, 1968; Strauss, 1966). There is yet, however, no statistical evidence to support the theorizing of these clinicians, who work largely within an existential–phenomenological framework that rarely generates statistical data.

Measurement of Obsessive Compulsiveness

Over the past two decades, several questionnaires and scales have been devised that purport to measure obsessional traits and characteristics (Allen & Tune, 1975; Beloff, 1957; Blum, 1949; Caine & Hawkins, 1963; Comrey, 1965; Cooper, 1970; Gottheil, 1965b; Grygier, 1956; Kline, 1969; Lazare, Klerman, & Armor, 1966, 1970; Sandler & Hazari, 1960). Kline (1969) argued that there has really not been an abundance of empirical research on obsessional personality, primarily because there is no fully accepted measure of obsessional traits or obsessional symptoms. Most, if not all, of the existing scales are not standardized, nor is there sufficient evidence for their reliability and validity to justify a rational choice of one over the other. According to Kline the major personality inventories, like the Sixteen Personality Factor Questionnaire (Cattell & Eber, 1957), the Eysenck Personality Inventory, the Maudsley Personality Inventory, and the Minnesota Multiphasic Personality Inventory (MMPI), do not contain a measure of obsessional traits and characteristics. The Psychasthenia scale of the MMPI, designated as Scale 7, is sometimes referred to as a measure of obsessive–compulsiveness; however, there is good reason to suspect that the Psychasthenia scale is more a general measure of classically neurotic concerns, preoccupations, and characteristics, namely, anxiety, withdrawal, immobilization, agitation, and so on, rather than a specific measure of obsessive–compulsive behavioral tendencies (e.g., Dahlstrom, Welsh, & Dahlstrom, 1972; Drake & Oetting, 1959). One of the more promising measures to date is the Lazare-Klerman Trait Scales (Lazare et al., 1966, 1970), a factor analytically derived instrument that purports to measure obsessional, hysterical, and oral dependent personality. This self-report inventory contains 140 true–false items that are scored into 20 trait scores, each based on seven items. These 20 traits are reported in four separate factor analytic studies to combine to three orthogonal factors that closely mirror obsessive, hysterical, and oral character traits, as described in the clinical literature (Lazare et al., 1966, 1970; Paykel &

Prusoff, 1973; Van Den Berg & Helstone, 1975).

Empirical Validation of Obsessive–Compulsive Personality

Clinical descriptions are highly consistent in terms of the defenses, traits, and behavioral styles thought to be defining characteristics of the obsessive–compulsive personality. The findings of statistical studies have in the majority of cases generally supported the descriptions of the anal trait clusters found in the clinical studies. The bulk of these studies have been of the correlational and factor analytic type (Barnes, 1952; Beloff, 1957; Brooks, 1969; Comrey, 1965; Cooper & Kelleher, 1973; Finney, 1961, 1963; Gottheil, 1965b; Gottheil & Stone, 1968; Hetherington & Brackbill, 1963; Kline, 1968; Lazare et al., 1966, 1970; Mandel, 1958; Paykel & Prusoff, 1973; Rapaport, 1955; Sandler & Hazari, 1960; Schlesinger, 1963; Sears, 1943; Stagner, Lawson, & Moffitt, 1955; Stagner & Moffitt, 1956; Stone & Gottheil, 1975; Van Den Berg & Helstone, 1975).

Finney (1961, 1963) and Beloff (1957) found the traits of obstinacy, parsimony, and orderliness to be correlationally related in the children they studied.

Barnes (1952), Stagner et al. (1955), and Stagner and Moffitt (1956) sought to show through factor analytic methods that traits related to the psychosexual stages delineated in Freudian theory were empirically grouped in adult subjects. The results of these early factor analyses must be viewed as equivocal and difficult to accurately interpret because of inconsistent definitions of psychosexual stages and traits and the dubious adequacy of the measures used (Gottheil & Stone, 1968). In the study by Barnes, a *meticulous* rather than an anal factor per se was identified using the responses of 266 male college students to lists of items composed for the study. However, grouped together on this factor were traits of orderliness, reliability, law abidance, and cleanliness, in addition to meticulousness. Rigidity, sadism, and defiant resentment loaded on another factor termed *externalized aggression*.

Other factor analytic investigations appear to be more consistent with theory and clinical description. Schlesinger (1963) performed a factor analysis of 154 items taken from various anality questionnaires and extracted 12 factors, namely, responsibility in dealing with others, regularity and meticulousness, retentiveness as a style of life, obstinacy, rigidity, frugality, concern about dirt and contamination, orderliness, self-righteous hostility and competitiveness, anxiety over possible loss of control, sensitivity to smells, and retentiveness in relation to possessions. Cooper and Kelleher (1973) carried out a principal-components analysis of the Leyton Obsessional Inventory (Cooper, 1970) using approximately 300 normal subjects divided by sex and nationality (Irish and English). Three distinct components were derived from four separate analyses and termed (a) concern with being clean and tidy, (b) feeling of incompleteness, (c) checking and repetition. The second component appeared to relate to a need for closure, as reflected in the following item: "Even when you have done something carefully, do you often feel that it is somehow not quite right or complete?" Other less distinct components were found, one of which suggested the label *methodical*.

In studies by Lazare et al. (1966, 1970), alluded to above, factor analysis was employed to explore three personality patterns derived from psychoanalytic theory: oral, obsessive, and hysterical. The three personality patterns were defined by approximately 20 traits obtained from a review of the clinical literature. Each of these traits was then reliably measured from groups of true and false items rated by samples of female psychiatric inpatients and outpatients. The traits so measured formed three clusters that corresponded quite closely to psychoanalytic descriptions of oral, obsessive, and hysterical personality patterns. In the earlier study (Lazare et al. 1966), all of the defining traits of the obsessive personality factor were correctly predicted from theory, for example, orderliness, parsimony, rejection of others, emotional constriction, obstinacy, severe superego, rigidity, and perseverance. One predicted obsessive trait, self-doubt, however, only had a factor loading of .12. In their later study, Lazare

et al. (1970) replicated their earlier findings with the exception of the trait of obstinacy. Virtually identical defining traits have been reported by Paykel and Prusoff (1973), who used the responses of male and female recovered depressives to the Lazare-Klerman Trait Scales, and by Van Den Berg and Helstone (1975), who used the Lazare-Klerman Trait Scales in Holland with samples of psychiatric and normal females.

Gottheil (1965a) sought to investigate the extent to which mental health professionals agree in their use of the terms *anal* and *oral* character. A group of psychiatrists and clinical psychologists ($N = 20$) completed questionnaires composed of items derived from clinical descriptions of oral and anal character types and then were asked to predict how a typical oral character would answer the oral questionnaire and how an anal character would typically answer a questionnaire on anal character traits. The degree of consistency demonstrated in the responses of the professional subjects was found to be highly significant statistically, suggesting that mental health experts possess similar conceptions of these character types. Significant agreement was also found among the subjects on a majority of the items that constituted the two questionnaires ($p < .03$). Whereas in both instances agreement was quite high, there was less agreement on the conceptualization of the oral character than of the anal character, suggesting that the concept of the latter has been more clearly and unambiguously described.

In another study, Gottheil (1965b) administered his newly constructed oral and anal questionnaires to 200 army enlisted men. Item analyses indicated that the various characteristics attributed by expe-t judges to the anal and oral character types are empirically associated in the responses of normal adult male subjects.

In one factor analytic study using the same sample of army recruits, Gottheil and Stone (1968) derived an oral and an anal trait factor from responses both to the items constituting the oral and anal trait questionnaires and to items concerned with mouth and bowel habits. Oral and anal subfactors were also derived that were quite consistent with psychoanalytic descriptions. The five anal subfactors identified were termed (a) rigidity; (b) obsessive rumination; (c) perfectionism (which includes orderliness, persistence, and a tendency to be critical); (d) parsimony, possessiveness, and checking and rechecking; and (e) a practical point of view. However, in the overall questionnaire factor analysis, the oral and anal trait factors were weak. Together they accounted for only 5.3% of the variance in the total set of items. Moreover, only 23% of the variance in the anal trait scale was accounted for by the five anal trait subfactors that were extracted. Thus, despite the kinds of items selected, neither anal nor oral character structure emerged as the strongest organizing factor. All of these results were for the most part confirmed in a later study (Stone & Gottheil, 1975) that factor analyzed the responses of samples of neurotic and psychosomatically involved outpatients to the same sets of items employed in the previous study.

Studies by Gottheil (1965b), Gottheil and Stone (1974), and Kline (1967) provide evidence that obsessional character traits are normally distributed in samples of normals and of neurotic and psychosomatic patients. Therefore, adult subjects that constitute both normal and clinical groups cannot be characterized as either having or not having obsessive tendencies, but rather must be characterized as having more or less of them.

Obsessional Symptoms Versus Obsessional Personality

Several studies have addressed the issue of whether obsessional symptoms can be reliably differentiated from obsessional personality traits and characteristics. As discussed in the Introduction, the distinction between personality and symptoms was emphasized by Freud and his contemporaries, as well as by virtually all later clinicians.

Using a sample of 100 neurotic patients, approximately equally divided between men and women, Sandler and Hazari (1960) factor analyzed responses to 40 items related to obsessive–compulsive character traits and symptoms and found two relatively inde-

pendent, orthogonal personality constellations or dimensions quite similar to the clinical distinctions between obsessional character traits and obsessional neurotic symptomatology. The items that loaded on the character trait dimension present a picture of an exceedingly systematic and methodical person who likes a well-ordered life, is consistent and punctual, and is meticulous in his or her use of words. The individual dislikes half-done tasks and finds interruptions in plans and goal-directed activity irksome. He or she pays much attention to detail and has a strong aversion to dirt. These characteristics are well integrated in the personality, that is, they are ego syntonic and are frequently viewed as a source of pride and esteem by their possessors.

The other dimension Sandler and Hazari identified corresponded well to descriptions of obsessional neurotic symptomatology, for example, when life is severely disrupted by the intrusion of unwanted thoughts and compulsive acts and by worry, doubt, and procrastination. Unlike the obsessional traits, these unwanted thoughts and impulses were experienced as alien and disturbing. In psychoanalytic terms, the traits represent a successful or adaptive ego defense, whereas the symptoms are evidence of a breakdown in defense mechanisms (Kline, 1967). Sandler and Hazari concluded that although the two dimensions were orthogonal, this does not necessarily mean that in many instances subjects will not show a mixture of both dimensions; nor does it necessarily imply that both do not share common dynamics and etiology, for example, conflicts over anal urges. Subsequent studies by Foulds, Caine, Adams, and Owen (1965), Kline (1967), and Meares (1971) found evidence to support Sandler and Hazari's original distinction between ego-syntonic traits and ego-dystonic (or alien) symptomatology. Slade (1974) similarly concluded that factor analytic investigations in which both obsessional-trait and obsessional-symptom items have been included strongly suggest the existence of separate trait and symptom factors. Whether a single trait and a single factor emerge or a number of both factors emerge may be primarily dependent on the range of behavior studied.

The results of two studies (Ingram, 1961b; C. M. Rosenberg, 1967) suggest that there is some relationship between obsessional neurosis and obsessional personality. Rosenberg investigated the personality characteristics of 47 obsessional neurotics by means of psychiatric ratings and performance on selected personality inventories. Of this sample, 25 were judged to have an obsessional premorbid personality, which was quite congruent with clinical descriptions, that is, orderly, rigid, obstinate, dependable, pedantic, and so forth. Psychometric studies like those of Sandler and Hazari (1960), however, do suggest the independence of obsessional illness and obsessional character. Moreover, clinical observation suggests that at least some obsessional neurotics never could be said to have had a premorbid obsessional character makeup (e.g., Rack, 1977). In addition, as Paykel and Prusoff (1973) pointed out, obsessional patients with a corresponding premorbid obsessional makeup represent a small and not necessarily typical segment of individuals with obsessive–compulsive personalities. Clearly there is no necessary one-to-one relationship between obsessional personality and obsessional neurosis, despite the occasional finding that more obsessive–compulsive neurotics than would be expected by chance show evidence of a premorbid obsessional personality.

Several studies have been specifically concerned with the relationship between obsessional traits and the personality dimension of neuroticism. The findings of factor analytic studies such as those of Sandler and Hazari suggest that obsessional traits may relate inversely to neuroticism, whereas obsessional symptoms may relate positively to measures of neuroticism, maladjustment, and emotional instability. With one exception (Orme, 1965), the few studies that have addressed this issue report these relationships to exist (Kline, 1968; Meares, 1971; Paykel & Prusoff, 1973). Kline (1967) factor analyzed the Sandler-Hazari Scale (1960), the Beloff Anal Test (1957), and the MMPI, using the responses of a normal sample of 93 college students. Three relevant factors emerged: general emotional instability, a factor of obsessional character traits, and a factor of so-

cial introversion. Neither the Beloff nor the Sandler-Hazari measure (both measuring obsessional character traits) loaded highly on the emotional instability factor, while the Sandler-Hazari measure of obsessional symptoms loaded highly on both the social introversion and emotional instability factors. In a subsequent study, Kline (1968) factor analyzed the MMPI, the Beloff Anal Test, the Sandler-Hazari Scale, and his newly devised anal scale (AI3), a measure of obsessional traits, and discovered that his measure also did not load on the emotional instability factor that runs through the clinical scales of the MMPI. Meares, using a sample of 32 patients with spasmodic torticollis, found a moderate positive correlation ($r = .6$, $p <$ .001) between a measure of neuroticism (the Eysenck Personality Inventory) and the obsessional symptoms section of the Sandler-Hazari Scale. Obsessional traits also measured by the Sandler-Hazari Scale were negatively related to the Eysenck Personality Inventory Neuroticism scale ($r = -.31$). Orme (1965) reported that obsessional character traits correlated moderately with a measure of emotional instability, Cattell's (Cattell & Eber, 1957) 13-item 'O' Factor Scale, in samples of normals and obsessional neurotics, but, as Paykel and Prusoff pointed out, Orme's measure of traits was derived from Sandler and Hazari's second dimension of ego-alien phenomena and relates more to obsessional symptoms than to obsessional traits. Paykel and Prusoff reported additional evidence that neuroticism, as measured by the Maudsley Personality Inventory, relates inversely to obsessional traits, as measured by the Lazare-Klerman Trait Scales ($r = -.23$, $p < .05$), in a sample of 131 recovered depressed male and female psychiatric patients. In a recent study, however, Pollak (1978), using the Lazare-Klerman Trait Scales as a measure of obsessive personality, found in a sample of graduate students ($N = 114$) that obsessive personality correlated negatively with several self-actualization variables of the Personal Orientation Inventory (Shostrom, 1966; r ranged from $-.16$ to $-.40$). If one views this inventory as a measure of optimal emotional functioning or positive mental health, then the inverse relationships found at least suggest that obsessive personality may reflect a somewhat immature, if not neurotic, character structure.

In summary, with some notable exceptions, studies indicate significant positive relationships between measures of obsessional neurosis and measures of emotional instability and significant negative relationships between obsessional trait measures and emotional instability (Slade, 1974).

Obsessive–Compulsive Personality and Introversion–Extraversion

According to the personality model of Eysenck (1947, 1959, 1960), two orthogonal dimensions are used to account for the psychoneuroses, namely, neuroticism and extraversion–introversion. Hysterical character disorders are viewed, for example, as disturbances of the neurotic extravert, whereas obsessive–compulsive disorders are classified as disorders of the neurotic introvert. Thus, according to this paradigm, obsessive–compulsive and hysterical patients are conceptualized as occupying opposite ends of a neurotic, introversion–extraversion continuum.

C. M. Rosenberg (1967) found, in a sample of obsessional neurotics, significantly lower than average scores both on the Extraversion scale of the Maudsley Personality Inventory and on the second-order extraversion factor of the Sixteen Personality Factor Questionnaire (Cattell & Eber, 1957). Several other studies (e.g., Barret, Caldbeck-Meenan, & White, 1966; Caine & Hope, 1964; Forbes, 1969; Foulds et al., 1965) have found sizable correlations in the .70 to .80 range between performance on the Hysteroid-Obsessoid Questionnaire (Caine & Hawkins, 1963) and performance on the Extraversion scale of the Maudsley Personality Inventory and the second-order factor Extraversion scale of Cattell's Sixteen Personality Factor Questionnaire in psychiatric and nonpsychiatric samples, with subjects classified as obsessoid consistently scoring significantly lower on the extraversion measures than subjects classified as hysteroid. The results of these studies suggest that obsessive and hysterical personalities can be conceptualized as opposite ex-

tremes on an introversion–extraversion continuum, with obsessional personality highly related to introversion and hysterical personality to extraversion.

Paykel and Prusoff (1973), however, could not find any relationship between obsessional traits, as measured by the Lazare-Klerman Trait Scales, and the introversion–extraversion dimension, as measured by the Extraversion scale of the Maudsley Personality Inventory, in a sample of recovered depressed patients ($N = 131$). Moreover, in some earlier factor analytic research cited above, Kline (1967) found that obsessional character traits measured by both the Sandler-Hazari Scale of obsessional traits and symptoms and the Beloff Anal Test did not load highly on a social-introversion factor, whereas the Sandler-Hazari measure of obsessional symptoms, with a loading of .512, did.

It is interesting that when measured by the Hysteroid-Obsessoid Questionnaire, obsessive–compulsive personality appears positively related to introversion, but that when measured with other instruments like the Lazare-Klerman Trait Scales, it seems independent of the introversion–extraversion personality dimension.

Obsessive–Compulsive Personality and Response to Experimental Tasks

A number of studies have attempted to discover whether obsessive–compulsive personalities respond in experimental situations in ways congruent with predictions made from theory and clinical observation. For example, is there empirical evidence for the notion that obsessive individuals crave order and structure and strive to be methodical and efficient in their actions?

B. G. Rosenberg (1953) compared the performance of psychotherapy patients with pronounced obsessive–compulsive tendencies with a normal control group on a visual memory task that involved choosing from a multiple-choice format the ambiguous design previously seen in tachistoscopic presentation. The alternative choices varied in terms of degree of symmetry, and as expected, he found that obsessive–compulsive subjects tended to favor the more symmetrical choices. This was interpreted as reflecting a particular need to impose order, uniformity, and congruity on visual perception. It is an interesting finding that has never been directly cross-validated. Studies of a somewhat related nature, however, that used measures such as the Breskin nonverbal test of rigidity to explore correlates of perceptual rigidity have been reported (e.g., Breskin, 1968; Breskin, Gorman, & Hochman, 1970; Primavera, Simon, & Hochman, 1974).

In a more recent study that used a sample of 40 college men, Rosenwald (1972) found significant relationships between a measure of anxieties reflective of anal concerns and time spent bringing order to a disorderly situation, that is, straightening up of a pile of scattered magazines. No relationship, however, was found between time spent reorganizing the magazines and another questionnaire measure that specifically concerned anxiety about dealing with dirt. No relationship could be found, as well, between time spent on the reorganization task and efficiency in identifying geometric forms with their hands while the objects were immersed in a dirty, malodorous, feces-like medium. Poor or inefficient performance on this latter task was interpreted as indicative of weak defenses against anal impulses.

There are a few studies to date that have investigated differences in degree of anal orientation as a function of occupational choice. Would, for example, more obsessional individuals be found in occupational pursuits that emphasize being orderly and methodical? Using a Q-sort technique, Weinstein (1953) found, as predicted, that engineering students were more anal retentive than law or social work students. Using projective measures like the Rorschach test and the Bender-Gestalt test, Segal (1961) found partial support for his hypothesis that accounting students would demonstrate a greater degree of anal orientation than would creative writing students. He found, as expected, that the accounting group was less tolerant of ambiguity, more restrained in expression of hostile affect, and generally more emotionally controlled. No differences were found between the groups, though, in their use of compulsive defenses

and conformity to social expectations. Consistent with her predictions, Schlesinger (1963) found that accounting and engineering students exhibited a significantly greater anal orientation than did educational psychology students. White (1963), in a study of experienced clerical bank personnel, found an intense dislike of dirt and disorder not found to the same degree in a control group sample.

Although these studies of vocational activity in the main support psychoanalytic theorizing and clinical description insofar as they suggest positive relationships between obsessional characteristics and involvement in the more compulsive kinds of vocational pursuits, the findings would be strengthened if comparable results were to be found in studies that include only individuals actually practicing in different professions. With the exception of the study by White, findings have been based solely on responses of students and not practitioners.

A few studies have explored relationships between anal characteristics and verbal recall ability, a dependent variable that appears to bear some relation to both the presumed orderliness and retentiveness of obsessive–compulsive individuals. Adelson and Redmond (1958) hypothesized that anal retentive individuals should possess methodical and efficient ways to process and retain information. Using the performance of a sample of 61 college women on the Blacky Pictures (Blum, 1949), they found, as predicted, that anal retentive subjects showed significantly greater ability for immediate and delayed recall of prose passages than did anal expulsive subjects or subjects classified as neutral.

Pedersen and Marlowe (1950), in an earlier study, were unable to duplicate these findings using a male sample, and Marcus (1963), using the performance of female college students on the Blacky Pictures, found anal retentives to be superior to anal expulsives only with delayed as opposed to immediate recall of words. D. F. Fisher and Keen (1972) used the Blacky Pictures with a group of army men and found no significant relationships between the anality measures of retentiveness and expulsiveness and measures of recall of verbal material. As S. Fisher and Greenberg (1977) pointed out, results are mixed, with findings more congruent with theory in studies that used female subjects.

Several studies have investigated the presumed tendency of obsessive–compulsive personalities to be obstinate and negativistic. In a series of studies, it was found that, unlike "oral" subjects, "anal" subjects were extremely difficult to verbally condition using positive reinforcement (Cooperman & Child, 1971; Noblin, Timmons, & Kael, 1966; Timmons & Noblin, 1963). In a similar kind of investigation, Tribich and Messer (1974) found that oral subjects reported an amount of light movement similar to that reported by an authority figure when viewing the autokinetic phenomenon in their presence, whereas anal subjects tended to respond in opposition to the authority. In all four of the experimental studies that used the Blacky Pictures as a measure of orality and anality and that used college students as subjects, results were consistent with theory and the clinical impression that anal characters are often negativistic and resistant to control from authority figures.

The results of other studies suggest positive relationships between measures of anal character traits and such predicted oppositional behavioral features as nonacquiescent response set (Couch & Keniston, 1960), intense dislike for a task engaged in under conditions of forced compliance (Bishop, 1967), and resistance to attitude change (Rosenwald, 1972).

In obsessional personalities the trait of obstinacy is often linked with rigid, defiant, and hostile attitudes. In fact, the presumed rigidity of the obsessive–compulsive is thought to manifest itself in most, if not all, aspects of the individual's behavior, for example, in thought processes, perceptual style, verbal expression, motor activity, and so on (e.g., Reich, 1949; Shapiro, 1965). Authoritarianism has traditionally been linked in personality research literature to variables such as rigidity, low tolerance for ambiguity, and aggressive and hostile attitudes (e.g., Adorno, Frenkel-Brunswick, Levinson, & Sanford, 1950). Three studies that related obsessive–compulsive characteristics to rigid authoritarian beliefs and attitudes found positive relationships between the two, as might be

expected (Centers, 1969; Farber, 1955; Rogers & Wright, 1975). Farber found a correlation of .37 ($p < .01$) between the traits of orderliness, frugality, and obstinacy and what he termed political attitudes of *aggressive conventionality* (which emphasize strong antipathy to communism) in a fairly large sample of male and female college students ($N = 130$). Although Rabinowitz (1957) could not replicate Farber's results using a similar sample, some years later Centers (1969) found a relationship of comparable magnitude between a measure of anality and hard-line conservative attitudes on such issues as welfare and law and order in a cross-sectional sample of over 500 adults. Along similar lines, Rogers and Wright (1975) reported a sizable correlation of .62 ($p < .01$) between authoritarianism as measured by the California F Scale and obsessive–compulsiveness as measured by the Psychasthenia scale of the MMPI in a mixed sample of 38 undergraduates.

The character trait of obstinacy in obsessive–compulsive personalities has also been associated with conflictual and ambivalent attitudes toward the recognition and expression of hostile feelings. There are only a few studies that have experimentally explored this issue. In the study by Segal (1961), alluded to earlier, the presumably more anal student accountant group demonstrated more restraint in the expression of hostile affect on projective measures than did a group of creative writing students. Gordon (1966, 1967) related the nature and types of psychological interpretations made by clinical psychologists and clinical psychology trainees to the personality dimension of anality as measured by the Grygier Anality Scales (Grygier, 1956). Consistent with the point of view of Fenichel (1945) and Schafer (1954) that anal character types often lack confidence, are indecisive, sometimes demonstrate strong reaction formations against hostility, and tend toward generalization in their thinking, she found that high-anal clinicians had less confidence in their interpretations, made fewer specific predictions, and identified less psychopathology in the case and test material presented to them than did clinician groups designated as low anal or neutral on the basis of performance on the Grygier Anality Scales. In the aforementioned study by Rosenwald (1972), which employed both questionnaire and behavioral measures of anality, some evidence was found to support predicted relationships between anal character orientation and anxiety-ridden, conflictual, and ambivalent attitudes toward hostile and aggressive feelings and actions, as measured by questionnaires and responses to specific action-oriented tasks. The results overall were mixed, however, in relating the measures of anality to the measures of hostility and aggression. In support of theory, though, Rosenwald found that subjects high on two questionnaire measures of anal anxiety demonstrated more anxiety and were slower in performing a doll-destruction task than were subjects low on the anality measures.

With regard specifically to the anal feature of indecisiveness that was explored in Gordon's research, Rosenwald, Mendelsohn, Fontana, and Portz (1966) compared the performance of male college students on a task that involved identification of geometric forms with the hands under two conditions. In one condition, hands were placed in a feces-like medium; in the other, the geometric forms were handled in water. Inefficient or blocked performance under the more unpleasant conditions was construed as indicative of anally linked anxieties and was found to be positively related to indecisiveness, as defined by the amount of time subjects thought they needed to make the necessary perceptual estimates.

Empirical support for the presumed anal character trait of parsimony was sought in a few studies (Lerner, 1961; Noblin, 1962; G. M. Rapaport, 1955; Rosenwald, 1972). According to classical analytic formulations, money is equated in the unconscious with the excretory product, and activities that involve hoarding, collecting, and preserving objects, especially those symbolic of the excretory product and anal function, become paramount in the behavioral style of the anal character, as a sublimation of the infantile wish to hold and retain feces.

In the study by Noblin, 60 hospitalized psychiatric patients placed into anal or oral groupings on the basis of responses to the

Blacky Pictures and on the basis of psychoanalytically formulated diagnoses were differentially rewarded by food or pennies in a verbal conditioning paradigm aimed at increasing the use of personal pronouns in constructing sentences. Anal subjects were found to be best motivated by the money reinforcer, whereas oral subjects were differentially responsive to the food reinforcer, as theory would predict. Using a sample of teenage boys ($N = 30$), Lerner (1961) reported that serious stamp collectors were either significantly more sensitive or selectively insensitive to anally tinged words than to neutral words presented on an audiotape in comparison to a matched group of boys with no collecting interests. Rosenwald (1972) found some evidence that degree of anal orientation was related to the wagering of less money in a gambling situation, but the relationship reported was only found with one of two questionnaire measures of anality and was not shown when the hand-immersion behavioral efficiency task, alluded to above, was employed as an index of anality. In an earlier study, Rapaport (1955) failed to find significant relationships between anality and degree of concern with money on the Thematic Apperception Test.

The character trait of parsimony should naturally carry over and be reflected in how the obsessive–compulsive individual approaches and manages time. Some early analytic formulations (e.g., Jones, 1918/1961) postulated that time also is an unconscious equivalent of the fecal product. Therefore, by virtue of their predominant anal fixations, obsessional personalities are considered to be particularly sensitive about wasting time and having to spend time against their will; they insist on being masters of their own time.

Three studies to date have investigated relationships between anal character traits and attitudes toward time (Campos, 1966; Gorman & Katz, 1971; Pettit, 1969). Campos found, in a sample of 100 male undergraduates, a positive relationship between anal retentive traits and the tendency to overestimate intervals of time. This finding is consistent with the view that time is overestimated precisely because it is something of special value to be retained if at all possible. Also reported was a positive relationship between degree of anality and the tendency to use time in a niggardly, thrifty, and cautious manner. In the study by Pettit, sizable positive correlations (.5 to .65; $p < .001$) were found between the Grygier Anality Scales (Grygier, 1956) and a *composite anality scale* devised for the study and a time questionnaire designed to measure the importance of time to the individual in ordering and controlling experience. Subjects used were 91 undergraduates. No significant sex differences were found. Gorman and Katz (1971) sought to replicate and extend Pettit's findings, using another sample of undergraduates ($N = 110$). They administered Pettit's time scale and composite anality scale, and in addition subjects completed four time scales that measure various time attitudes (Calabresi & Cohen, 1968). The time-anxiety scale measures uncomfortable feelings and thoughts about the future and a frustrated longing for the past. The time-submissiveness measure reflects dutiful and conforming attitudes toward time. Analysis of the data confirmed Pettit's findings of a strong relationship between anality and time attitudes, but only for specific aspects of time attitudes. Significant relationships were found between all time measures and the anality measure, except for a nonsignificant and quite low correlation ($r = .14$) between anality and time possessiveness. A correlation of greater magnitude was expected, given the presumed strong retentive orientation of obsessive individuals. The overall pattern of the results did suggest that the constellations of time attitudes and anal character traits might be more fruitfully conceptualized under the rubric of *rigidity* or *obsessive–compulsive character style,* a category which does not reflect psychopathological behavior, but simply reflects a distinct personality style.

Directions for Further Research

More study of the perceptual correlates of obsessive–compulsive personality is urged. Although there have been some intriguing findings (e.g., B. G. Rosenberg, 1953), very little research has been conducted in this

area to date. What, for example, would the relationship be between obsessive–compulsive personality and the field dependence–independence construct of Witkin and his associates (Witkin et al., 1954).

More research on the relationship of obsessive–compulsive personality to measures of aesthetic sensitivity and indices of creative thinking in verbal and nonverbal mediums is also recommended.

The particular kinds of meanings ascribed by obsessive–compulsive personalities to critical life situations and tasks, for example, vocational choice, marriage, death and dying, and so on, would also be an interesting area to explore. Is there, for example, a difference in the way obsessional personalities conceptualize or personify death and dying, as opposed to the perceptions of individuals considerably less obsessional in their orientation to life or of individuals with fundamentally different personality styles, for instance, those designated as oral, hysterical, impulsive, and so forth? If differences are found, are they, in fact, consistent with clinical observation and predictions from theory? A study of meaning would begin to focus more on how obsessional personalities experience the world and might produce empirical data that would help to better evaluate the existential, phenomenological points of view on obsessional personality style (e.g., Becker, 1974; M. H. Miller & Chotlos, 1960; Strauss, 1966).

Conclusions and Summary

One of the difficulties in attempting to evaluate the findings of the empirically based research is that a diverse number of specific measures and types of measurement approaches have been employed. Many of the indices of anality and obsessive–compulsive characteristics that have been used possess questionable psychometric adequacy. For example, for a time, particularly in the 1950s and 1960s, the Blacky Pictures (Blum, 1949) was often used as a measure of anality; but the Blacky Pictures is a projective test that carries with it all of the problems of standardization, reliability, and validity that have come to be associated with the use of projective tests in personality assessment and research (e.g., Anastasi, 1968).

In addition, there have been few attempts to cross-validate some of the more intriguing findings, and some of the attempts to replicate have resulted in inconsistent findings. Moreover, only a few researchers have devoted themselves to repeated and in-depth treatment of a particular area of study in an attempt to refine measurement techniques and more closely study relationships suggested by theory or prior experimentation. On the basis of this review of the empirical literature on obsessive–compulsive personality, the following conclusions appear to be warranted.

1. Obsessive–compulsive personality, as a cluster of traits, appears to possess considerable empirical validity and to fairly closely adhere to clinical descriptions and predictions. This is true despite the fact that an array of measurement approaches and specific measurement instruments have been employed in an attempt to correlate measures of anality with various behavioral indices.

2. Obsessive–compulsive personality can be statistically differentiated from obsessive–compulsive symptomatology through factor analysis.

3. In most instances obsessive–compulsive personality does not appear to be positively related to measures of neuroticism, whereas obsessional symptoms do; however, findings are somewhat inconsistent on this issue.

4. Obsessive–compulsive personality may be independent of an introversion–extraversion classification scheme, but more study with diverse measures of obsessive–compulsive personality is needed to help clarify this issue.

5. Obsessive–compulsive traits appear to be normally distributed.

6. There appears to be little evidence in favor of classical psychoanalytic theories about the psychogenesis of obsessive–compulsive personality. Overall, there is yet no strong empirical support for any etiologic explanation. There are, though, suggestions in clinical observation and in some statistically based research that obsessive–compulsive individuals often are the progeny of obsessive–compulsive parents and that obsessional or anal character structure develops, at least in part, from early learning, but not necessarily

exclusively out of the toilet training period per se. Rigid, compulsive parenting, however, may very well be maximal before and during the time that sphincter control is still in the process of development.

References

Abraham, K. [Contributions to the theory of the anal character.] In D. Bryan & A. Strachey (trans.), *Selected papers*. New York: Basic Books, 1953. (Originally published, 1921.)

Adelson, J., & Redmond, J. Personality differences in the capacity for verbal recall. *Journal of Abnormal and Social Psychology*, 1958, *57*, 244–248.

Adorno, T. W., Frenkel-Brunswick, E., Levinson, D. J., & Sanford, R. N. *The authoritarian personality*. New York: Harper, 1950.

Allen, J. J., & Tune, G. S. The Lynfield obsessional compulsive questionnaires. *Scottish Medical Journal*, 1975, *20*, 21.

Anastasi, A. *Psychological Testing*. London: Macmillan, 1968.

Barnes, C. A. Statistical study of the Freudian theory of levels of psychosexual development. *Genetic Psychology Monographs*, 1952, *45*, 109–174.

Barret, W., Caldbeck-Meenan, J., & White, J. G. Questionnaire measures and psychiatrist ratings of a personality dimension. *British Journal of Psychiatry*, 1966, *112*, 413–415.

Becker, E. *The denial of death*. New York: Free Press, 1974.

Beloff, H. The structure and origin of the anal character. *Genetic Psychology Monographs*, 1957, *55*, 141–172.

Bernstein, A. Some relations between techniques of feeding and training during infancy and certain behavior in childhood. *Genetic Psychology Monographs*, 1955, *51*, 3–44.

Bishop, F. V. The anal character: A rebel in the dissonance family. *Journal of Personality and Social Psychology*, 1967, *6*, 23–36.

Blum, G. S. A study of the psychoanalytic theory of psychosexual development. *Genetic Psychology Monographs*, 1949, *39*, 3–99.

Breskin, S. Measurement of rigidity: A nonverbal test. *Perceptual and Motor Skills*, 1968, *27*, 1203–1206.

Breskin, S., Gorman, B., & Hochman, S. H. Nonverbal rigidity and perseveration. *Journal of Psychology*, 1970, *75*, 239–242.

Brooks, J. The insecure personality: A factor analytic study. *British Journal of Medical Psychology*, 1969, *42*, 395–403.

Caine, T. M., & Hawkins, L. G. Questionnaire measure of the hysteroid/obsessoid component of personality. *Journal of Consulting Psychology*, 1963, *27*, 206–209.

Caine, T. M., & Hope, K. Validation of the Maudsley Personality Inventory E scale. *British Journal of Psychology*, 1964, *55*, 447–452.

Calabresi, R., & Cohen, J. Personality and time attitudes. *Journal of Abnormal Psychology*, 1968, *73*, 431–439.

Campos, L. P. Relationship between time estimation and retentive personality traits. *Perceptual and Motor Skills*, 1966, *23*, 59–62.

Carr, A. T. Compulsive neurosis: A review of the literature. *Psychological Bulletin*, 1974, *81*, 311–318.

Cattell, R. B., & Eber, H. W. *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign, Ill.: Institute for Personality and Ability Testing, 1957.

Centers, R. The anal character and social severity in attitudes. *Journal of Projective Techniques and Personality Assessment*, 1969, *33*, 501–506.

Comrey, A. L. Scales for measuring compulsion, hostility, neuroticism, and shyness. *Psychological Reports*, 1965, *16*, 697–700.

Cooper, J. Leyton Obsessional Inventory. *Psychological Medicine*, 1970, *1*, 48–64.

Cooper, J., & Kelleher, M. J. Leyton Obsessional Inventory: A principal components analysis on normal subjects. *Psychological Medicine*, 1973, *3*, 204–208.

Cooperman, M., & Child, I. L. Differential effects of positive and negative reinforcement on two analytic character types. *Journal of Consulting and Clinical Psychology*, 1971, *37*, 57–59.

Couch, A., & Keniston, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, 1960, *60*, 151–174.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook*. Minneapolis: University of Minnesota Press, 1972.

Drake, L. E., & Oetting, E. R. *An MMPI cookbook for counselors*. Minneapolis: University of Minnesota Press, 1959.

Durrett, M. E. The relationship of early infant regulation and later behavior in play interviews. *Child Development*, 1959, *30*, 211–216.

Eysenck, H. J. *Dimensions of personality*. London: Routledge & Kegan Paul, 1947.

Eysenck, H. J. *The manual of the Maudsley Personality Inventory*. London: University of London Press, 1959.

Eysenck, H. J. *Experiments in personality*. London: Routledge & Kegan Paul, 1960.

Farber, M. J. The anal character and political aggression. *Journal of Abnormal and Social Psychology*, 1955, *51*, 486–489.

Fenichel, O. *The psychoanalytic theory of neurosis*. New York: Norton, 1945.

Finney, J. C. The MMPI as a measure of character structure as revealed by factor analysis. *Journal of Consulting Psychology*, 1961, *25*, 327–336.

Finney, J. C. Maternal influences on anal or compulsive character in children. *Journal of Genetic Psychology*, 1963, *103*, 351–367.

Fisher, D. F., & Keen, S. L. Verbal recall as a function of personality characteristic. *Journal of Genetic Psychology*, 1972, *120*, 83–92.

Fisher, S., & Greenberg, R. P. *The scientific credibility of Freud's theories and therapy*. New York: Basic Books, 1977.

Forbes, A. R. The validity of the 16PF in the discrimination of the hysteroid and obsessional personality. *British Journal of Social and Clinical Psychology*, 1969, *8*, 152–159.

Foulds, G. A., Caine, T. M., Adams, A., & Owen, A. *Personality and personal illness*. London: Tavistock, 1965.

Freud, S. Character and anal eroticism. In P. Rieff (Ed.), *Collected papers* (Vol. 10). New York: Collier, 1963. (Originally published, 1908.)

Gordon, C. M. Some effects of information, situation, and personality on decision making in a clinical setting. *Journal of Consulting Psychology*, 1966, *30*, 219–224.

Gordon, C. M. Some effects of clinician and patient personality on decision making in a clinical setting. *Journal of Consulting Psychology*, 1967, *31*, 477–480.

Gorer, G. Themes in Japanese culture. *Transactions of the New York Academy of Sciences*, 1943, *5*, 106–124.

Gorman, B. S., & Katz, G. Temporal orientation of anality. *Proceedings of the 79th Annual Convention of the American Psychological Association*, 1971, *6*, 367–368. (Summary)

Gottheil, E. Conceptions of orality and anality. *Journal of Nervous and Mental Disease*, 1965, *141*, 155–160. (a)

Gottheil, E. An empirical analysis of orality and anality. *Journal of Nervous and Mental Disease*, 1965, *141*, 308–317. (b)

Gottheil, E., & Stone, G. C. Factor analytic study of orality and anality. *Journal of Nervous and Mental Disease*, 1968, *146*, 1–17.

Gottheil, E., & Stone, G. C. Psychosomatic aspects of orality and anality. *Journal of Nervous and Mental Disease*, 1974, *159*, 182–190.

Grygier, T. C. *Dynamic Personality Inventory*. London: National Foundation for Educational Research in England, 1956.

Hays, P. Determination of the obsessional personality. *American Journal of Psychiatry*, 1972, *129*, 217–219.

Hetherington, E. M., & Brackbill, Y. Etiology and covariation of obstinacy, orderliness, and parsimony in young children. *Child Development*, 1963, *34*, 919–943.

Holway, A. R. Early self-regulation of infants and later behavior in play interviews. *American Journal of Orthopsychiatry*, 1949, *19*, 612–623.

Honigmann, J. J. *Culture in personality*. New York: Harper & Row, 1967.

Huschka, M. The child's response to coercive bowel training. *Psychosomatic Medicine*, 1942, *4*, 301–328.

Ingram, I. M. Obsessional personality and anal-erotic character. *Journal of Mental Science*, 1961, *107*, 1035–1042. (a)

Ingram, I. M. The obsessional personality and obsessional illness. *American Journal of Psychiatry*, 1961, *117*, 1016–1019. (b)

Jones, E. Anal erotic character traits. In E. Jones, *Papers on psychoanalysis*. Boston: Beacon Press, 1961. (Originally published, 1918.)

Kline, P. Obsessional traits and emotional instability in a normal population. *British Journal of Medical Psychology*, 1967, *40*, 153–157.

Kline, P. Obsessional traits, obsessional symptoms, and anal eroticism. *British Journal of Medical Psychology*, 1968, *41*, 299–305.

Kline, P. The anal character: A cross-cultural study in Ghana. *British Journal of Social and Clinical Psychology*, 1969, *8*, 201–210.

Lazare, A., Klerman, G. L., & Armor, D. J. Oral, obsessive, and hysterical personality patterns. *Archives of General Psychiatry*, 1966, *14*, 624–630.

Lazare, A., Klerman, G. L., & Armor, D. J. Oral, obsessive, and hysterical personality patterns: Replication of factor analysis in an independent sample. *Journal of Psychiatric Research*, 1970, *7*, 275–279.

Lerner, B. *Auditory and visual thresholds for the perception of words of anal connotation: An evaluation of the "sublimation hypothesis" on philatelists*. Unpublished doctoral dissertation, Yeshiva University, 1961.

Mandel, H. *A Q-methology investigation of the oral and anal character as described by psychoanalytic theory*. Unpublished doctoral dissertation, New York University, 1958.

Marcus, M. M. *The relation of personality structure to the capacity for memory retention*. Unpublished doctoral dissertation, University of Pittsburgh, 1963.

Meares, R. Obsessionality, the Sandler-Hazari Scale, and spasmodic torticollis. *British Journal of Medical Psychology*, 1971, *44*, 181–182.

Miller, D. R., & Swanson, G. E. *Inner conflict and defense*. New York: Schocken, 1966.

Miller, M. H., & Chotlos, J. W. Obsessive and hysterical syndromes in the light of existential consideration. *Journal of Existential Psychiatry*, 1960, *1*, 315–329.

Noblin, C. D. *Experimental analysis of psychoanalytic character types through the operant conditioning of verbal responses*. Unpublished doctoral dissertation, Louisiana State University, 1962.

Noblin, C. D., Timmons, E. O., & Kael, H. C. Differential effects of positive and negative verbal reinforcement on psychoanalytic character types. *Journal of Personality and Social Psychology*, 1966, *4*, 224–228.

O'Connor, N., & Franks, C. M. Childhood upbringing and other environmental factors. In H. J. Eysenck (Ed.), *Handbook of abnormal psychology*. London: Pitman, 1960.

Orlansky, H. Infant care and personality. *Psychological Bulletin*, 1949, *46*, 1–48.

Orme, J. E. The relationship of obsessional traits to general emotional instability. *British Journal of Medical Psychology*, 1965, *38*, 269–271.

Paykel, E. S., & Prusoff, B. A. Relationships between personality dimensions: Neuroticism and extraversion against obsessive, hysterical, and oral per-

sonality. *British Journal of Social and Clinical Psychology,* 1973, *12,* 309–318.

Pedersen, F., & Marlowe, D. Capacity and motivational differences in verbal recall. *Journal of Clinical Psychology,* 1950, *16,* 219–222.

Pettit, T. F. Anality and time. *Journal of Consulting and Clinical Psychology,* 1969, *33,* 170–174.

Pollak, J. Relationships between psychoanalytic personality pattern, death anxiety, and self-actualization. *Perceptual and Motor Skills,* 1978, *46,* 846.

Primavera, L. H., Simon, W. E., & Hochman, S. H. Nonverbal rigidity and its relationship to performance on three standard reversible figures. *Journal of Psychology,* 1974, *86,* 61–63.

Rabinowitz, W. Anality, aggression, and acquiescence. *Journal of Abnormal and Social Psychology,* 1957, *54,* 140–142.

Rack, P. Clinical experience in the treatment of obsessional states. *Journal of International Medical Research,* 1977, *5,* 81–91.

Rado, S. Obsessive behavior. In S. Arieti (Ed.), *American handbook of psychiatry.* New York: Basic Books, 1959.

Rapaport, G. M. *A study of the psychoanalytic theory of the anal character.* Unpublished doctoral dissertation, Northwestern University, 1955.

Reich, W. *Character analysis.* New York: Orgone Institute Press, 1949.

Rogers, R., & Wright, E. W. Behavioral rigidity and its relationship to authoritarianism and obsessive-compulsiveness. *Perceptual and Motor Skills,* 1975, *40,* 802–804.

Rosenberg, B. G. Compulsiveness as a determinant in selected cognitive-perceptual performances. *Journal of Personality,* 1953, *21,* 509–516.

Rosenberg, C. M. Personality and obsessional neurosis. *British Journal of Psychiatry,* 1967, *113,* 471–477.

Rosenwald, G. C. Effectiveness of defenses against anal impulse arousal. *Journal of Consulting and Clinical Psychology,* 1972, *39,* 292–298.

Rosenwald, G. C., Mendelsohn, G. A., Fontana, A., & Portz, A. T. An action test of hypotheses concerning the anal personality. *Journal of Abnormal Psychology,* 1966, *71,* 304–309.

Salzman, L. *The obsessive personality.* New York: Science House, 1968.

Sandler, J., & Hazari, A. The "obsessional": On the psychological classification of obsessional character traits and symptoms. *British Journal of Medical Psychology,* 1960, *33,* 113–122.

Schafer, R. *Psychoanalytic interpretation in Rorschach testing.* New York: Grune & Stratton, 1954.

Schlesinger, V. J. *Anal personality traits and occupational choice: A study of accountants, chemical engineers, and educational psychologists.* Unpublished doctoral dissertation, University of Michigan, 1963.

Sears, R. R. *Survey of objective studies of psychoanalytic concepts* (Bulletin 51). New York: Social Science Research Council, 1943.

Sears, R. R., Rau, L., & Alpert, R. *Identification and child rearing.* Stanford, Calif.: Stanford University Press, 1965.

Segal, S. J. A psychoanalytic analysis of personality factors in vocational choice. *Journal of Counseling Psychology,* 1961, *8,* 202–210.

Sewell, W. H., Mussen, P. H., & Harris, C. W. Relationships among child training practices. *American Sociological Review,* 1955, *20,* 137–148.

Shapiro, D. *Neurotic styles.* New York: Basic Books, 1965.

Shostrom, E. L. *Manual for the Personal Orientation Inventory.* San Diego, Calif.: Educational and Industrial Testing Service, 1966.

Slade, P. D. Psychometric studies of obsessional illness and obsessional personality. In H. R. Beech (Ed.), *Obsessional states.* London: Methuen, 1974.

Stagner, R., Lawson, E. D., & Moffitt, J. W. The Krout Personal Preference Scale: A factor-analytic study. *Journal of Clinical Psychology,* 1955, *11,* 103–113.

Stagner, R., & Moffitt, J. W. A statistical study of Freud's theory of personality types. *Journal of Clinical Psychology,* 1956, *12,* 72–74.

Stone, G. C., & Gottheil, E. Factor analysis of orality and anality in selected patient groups. *Journal of Nervous and Mental Disease,* 1975, *160,* 311–323.

Straus, M. A. Anal and oral frustration in relation to Sinhalese personality. *Sociometry,* 1957, *20,* 21–31.

Strauss, E. *Phenomenological psychology.* New York: Basic Books, 1966.

Templer, D. Obsessive-compulsive neurosis: A review of research. *Comprehensive Psychiatry,* 1972, *13,* 375–383.

Timmons, E. O., & Noblin, C. D. The differential performance of anals and orals in a verbal conditioning paradigm. *Journal of Consulting Psychology,* 1963, *27,* 383–386.

Tribich, D., & Messer, S. Psychoanalytic character type and status of authority as determiners of suggestibility. *Journal of Consulting and Clinical Psychology,* 1974, *42,* 842–848.

Van Den Berg, P. J., & Helstone, F. S. Oral, obsessive, and hysterical personality patterns: A Dutch replication. *Journal of Psychiatric Research,* 1975, *12,* 319–327.

Weinstein, M. S. *Personality and vocational choice: A comparison of the self-conceptions and ideal self-conceptions of students in three profesional schools.* Unpublished doctoral dissertation, Western Reserve University, 1953.

White, J. C. Cleanliness and successful bank clerical personnel—A brief. *Journal of Counseling Psychology,* 1963, *10,* 192.

Whiting, J. W., & Child, I. L. *Child training and personality: A cross-cultural study.* New Haven, Conn.: Yale University Press, 1953.

Witkin, H. A., et al. *Personality through perception.* New York: Harper, 1954.

Received September 7, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 242–249

Ridge Regression: Bonanza or Beguilement?

William W. Rozeboom
University of Alberta, Edmonton, Canada

Ridge regression is an intriguing new toy for statistical estimation theory. But it is just that—a *toy* which may someday evolve into a useful if limited tool but is still too fragile to do real work. Specifically, ridge regression can indeed improve upon the accuracy of traditional estimates of regression parameters if background circumstances are right. But if they are not right—and how to diagnose this remains obscure—ridge regression incurs a loss of estimational accuracy.

Estimation of regression coefficients by ridge regression (Price, 1977) as a way to reduce the very large sampling uncertainty that invests classically estimated regression coefficients when the estimation sample's predictor distribution approaches multicollinearity is an exciting new methodological prospect. But exciting new prospects have a regrettable proclivity for promise without fulfillment—which is to say, more gently, that exploration usually comes up empty despite our need to keep looking if we are ever to find any new goodies at all—and ridge regression is no great exception to this rule. As will be acknowledged, the concept behind ridge regression has some merit; but its application to practical problems of predictor multicollinearity is both conceptually misleading and of uncertain practical value.

Preliminaries

It is misleading to characterize ridge regression as a partial solution to predictor multicollinearity for the simple reason that the degree to which the accuracy of ridge regression may or may not improve upon that of the classical procedure (called *ordinary least squares* or OLS by Price) for estimating predictor coefficients has nothing to do with the severity of predictor intercorrelations and is in principle equally helpful—or unhelpful—when the predictors are fully independent of one another. This is because (a) our predictor variables $X = \langle x_1, \ldots, x_m \rangle$ can always be rotated into an orthogonal basis $F = \langle f_1, \ldots, f_m \rangle$ for X space by a linear transformation $F = X\mathbf{W}$, in which $\mathbf{W}$ is an $m \times m$ matrix of rotation coefficients; (b) the column vectors $\mathbf{b}_X$ and $\mathbf{b}_F$ of the criterion variable's true regression coefficients on X and F, respectively, stand in relation $\mathbf{b}_X = \mathbf{W}\mathbf{b}_F$; (c) almost any method for estimating $\mathbf{b}_X$, OLS, and its ridge-regression extension in particular can be construed as a method that first estimates $\mathbf{b}_F$ as $\hat{\mathbf{b}}_F$ and then converts the latter into its estimate $\hat{\mathbf{b}}_X$ of $\mathbf{b}_X$ by transformation $\hat{\mathbf{b}}_X = \mathbf{W}\hat{\mathbf{b}}_F$; and (d) for OLS and ridge regression, when $f_1, \ldots, f_m$ are orthogonal, the expected squared error $\mathrm{Exp}[(\hat{b}_{X_i} - b_{X_i})^2]$ of the ith term of $\hat{\mathbf{b}}_X$ is a weighted sum of the expected squared errors of $\hat{b}_{F1}, \ldots, \hat{b}_{Fm}$, the weight of $\mathrm{Exp}[(\hat{b}_{Fj} - b_{Fj})^2]$ being the square of the ijth element of $\mathbf{W}$.[1] Point d explains why predictor near-multicollinearity can ravage the accuracy of $\hat{\mathbf{b}}_X$, though I defer details until

Requests for reprints should be sent to William W. Rozeboom, Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6G 2E9.

[1] I have slid quickly by a number of technical points here. For one, we assume that none of the predictors is exactly a linear function of the others in the sample, albeit this degeneracy can be approached as closely as we please. The sampling model envisioned here and later is the one with fixed predictor covariances, that is, we consider the sampling distribution of $\mathbf{b}$ under repeated sampling with sample size and the predictor covariances held constant at the values found in our actual sample. And when ridge regression's $\mathbf{K}$-matrix parameter is taken to be a sample-dependent random variable rather than a constant, it is possible for point d to be true of ridge regression only approximately, depending on the details of $\mathbf{K}$'s selection.

later. It also shows that the expected squared error of $\hat{b}_{X_i}$ can be less for ridge regression than for OLS only when this is also true for some of the $\hat{b}_{F_j}$ for orthogonal predictors $f_1, \ldots, f_m$.

The Nature of Ridge Regression

Estimating the simultaneous regression of criterion variable y on several orthogonal predictors is equivalent, that is, yields the same numerical results, for both OLS and ridge regression to estimating y's regression on each predictor separately.[2] Let us, therefore, analyze the nature of ridge regression for a single predictor x.

According to the standard sampling-theoretic regression model, $y = a + bx + e$, in which a and b are scalar constants (a being of no further interest here) and e is a residual variable, statistically independent of x, whose mean and variance under repeated sampling are respectively zero and an unknown quantity σ_e^2. Let σ_x^2, c_{yx}, and c_{ex} be, respectively, the variance of x, the covariance of y with x, and the covariance of e with x in the observed size-N sample from which we hope to estimate b. (We assume $\sigma_x^2 > 0$; otherwise b is not well-defined.) Then $c_{yx} = b\sigma_x^2 + c_{ex}$; while given a fixed N and σ_x^2 (see Footnote 1) and fully random sampling on e, it is easy to prove that

$$\mathrm{Exp}(c_{ex}) = 0, \quad \mathrm{Exp}(c_{yx}) = b\sigma_x^2,$$

and

$$\mathrm{Exp}(c_{ex}^2) = N^{-1}\sigma_e^2\sigma_x^2.$$

For any estimate $\hat{b}$ of b, let $S(\hat{b})$ be the expected squared error of $\hat{b}$, that is,

$$S(\hat{b}) =_{\mathrm{def}} \mathrm{Exp}[(\hat{b} - b)^2],$$

while henceforth $\dot{b}$ and $\tilde{b}$ are specifically b's OLS and ridge-regression estimates, respectively. Classically, OLS estimate $\dot{b}$ is defined to minimize $S(\dot{b})$ under the constraint that $\dot{b}$ is unbiased, leading to the computational formula $\dot{b} = c_{yx}/\sigma_x^2$. It follows that

$$\dot{b} = (b\sigma_x^2 + c_{ex})/\sigma_x^2 = b + c_{ex}/\sigma_x^2,$$

whence

$$\mathrm{Exp}(\dot{b}) = b$$

and

$$S(\dot{b}) =_{\mathrm{def}} \mathrm{Exp}[(\dot{b} - b)^2] = \mathrm{Exp}(c_{ex}^2)/\sigma_x^2 = \sigma_e^2/N\sigma_x^2,$$

which says inter alia that $\dot{b}$ is unbiased and that its expected squared error is independent of the parameter being estimated. $S(\dot{b})$ is not, however, independent of the predictor variable's variance—nor can it be for any estimate of b, insomuch as b itself is dependent on σ_x^2. For if x is linearly rescaled to shift its standard deviation from σ_x to $\sigma_{x^*} = s\sigma_x$, b is correspondingly rescaled to $b^* = s^{-1}b$, leaving $b^2\sigma_x^2$ (the amount of y variance linearly accounted for by x) invariant. But as b is multiplicatively adjusted by scaling factors, so of necessity is any estimate $\hat{b}$ of b together with its expected squared error. In particular, $S(\hat{b})$ can be made arbitrarily large by making σ_x sufficiently small. The ratio of $S(\hat{b})$ to b^2, however, is generally independent of the predictor scale. For OLS estimation in particular,

$$S(\dot{b})/b^2 = \sigma_e^2/Nb^2\sigma_x^2 = (1 - \rho_{yx}^2)/N\rho_{yx}^2,$$

where ρ_{yx} is the correlation between y and x in the population sampled[3] and is approximated (with a positive bias that vanishes with increasing N) by the sample correlation.

Ridge regression seemingly aspires to improve upon OLS estimation by correcting for the very large $S(\dot{b})$ that accompanies very small σ_x^2. Specifically, the ridge-regression estimate of b is $\tilde{b} = c_{yx}/(\sigma_x^2 + k)$, in which k is a small positive quantity whose numerical value is the task of ridge-regression theory to provide. (More generally, for multiple pre-

---

[2] More precisely, this is true of orthogonal predictors' OLS-estimated regression coefficients. The estimated uncertainty associated with the estimate of each b_{X_i} is a mildly increasing function of the number of additional predictors, insomuch as estimating coefficients also for the latter incurs a degrees-of-freedom loss from the total sample size.

[3] More precisely, ρ_{yx} is the correlation between y and x in a very large population that randomly samples e while having the same predictor variance as the observed sample (see Footnote 1). Also, when x is only one of m orthogonal predictors, ρ_{yx} is not the zero-order correlation between y and x, but is their partial correlation after the other predictors have been partialed out.

dictors, ridge regression alters some or all roots of the predictor covariance matrix by small increments that in the method's simplest version are the same for all roots.) However, we have just noted that σ_x^2 can be set at any stipulated positive value by choice of predictor scale; hence if $S(\tilde{b})$ can be less than $S(\dot{b})$ for some k when σ_x^2 is very small, the same degree of improvement must be possible when σ_x^2 is arbitrarily large.

The essential character of ridge regression, and its relation to OLS estimation, is evident in the relation

$$\tilde{b} = c_{yx}/(\sigma_x^2 + k)$$
$$= (c_{yx}/\sigma_x^2)[\sigma_x^2/(\sigma_x^2 + k)] = \dot{b}h, \quad (1)$$

where $h =_{\text{def}} (1 + k/\sigma_x^2)^{-1}$ and $h < 1$ if $k > 0$. Thus $\tilde{b}$ is simply an attenuation of $\dot{b}$ by a to-be-selected shrinkage factor h (cf. Mayer & Wilke, 1973). Since

$$\tilde{b} - b = \dot{b}h - b = (\dot{b} - b)h + b(h - 1),$$

or

$$(\tilde{b} - b)^2 = (\dot{b} - b)^2h^2 + 2(\dot{b} - b)bh(h - 1) + b^2(h - 1)^2,$$

the expected squared error of b conditional on a fixed (i.e., sample-independent) value of h is

$$S[\tilde{b}_{(h)}] = h^2S(\dot{b}) + b^2(h - 1)^2$$
$$= S(\dot{b})\cdot[(1 - h_0)^{-1}(h - h_0)^2 + h_0], \quad (2)$$

where

$$h_0 =_{\text{def}} [1 + S(\dot{b})/b^2]^{-1}$$
$$= [1+(1 - \rho_{yx}^2)/N\rho_{yx}^2]^{-1}$$
$$= N\rho_{yx}^2/[(N - 1)\rho_{yx}^2 + 1]. \quad (3)$$

The choice of h that minimizes $S[\tilde{b}_{(h)}]$ is thus $h = h_0$, given which the inaccuracy of b compared with that of OLS estimate $\dot{b}$ is $S[\tilde{b}_{(h_0)}]/S(\dot{b}) = h_0$. Even if h is chosen non-optimally, moreover, $S[\tilde{b}_{(h)}]$ is still less than $S(\dot{b})$ if h differs from h_0 by less than $1 - h_0$. Hence if there is a practical way to choose h in this interval with suitably high probability, ridge regression will indeed be more accurate than OLS estimation even if the improvement cannot amount to much unless h_0 is appreciably less than unity.

Before considering whether this prospect can be realized, however, observe that nothing in Equations 1, 2, and 3, save replacement of $S(\dot{b})/b^2$ by its analysis in terms of ρ_{yx}, is specific to estimation of regression coefficients. The term $\dot{b}$ can be any unbiased estimator from which adjusted estimator $\tilde{b} = \dot{b}h$ is obtained by an attenuation factor h. If this correction procedure can be made to work for estimation of regression coefficients, then we must anticipate that it may also be workable for many other classic unbiased estimators, starting with the sample mean as an unbiased estimate of the population mean. Either the logic of ridge regression calls for wholesale reappraisal of all established statistical estimation procedures or there is something impractically fragile about its applicability.

There is good reason to suspect the latter. For, any procedure that endeavors to select a value of h as close to h_0 as possible is in effect a procedure for estimating h_0; and since h_0 is so importantly a function of b, this is tantamount to deriving $\hat{h}_0$ from an estimate of b, the very parameter that is at issue in the first place. Although ridge regression's logic is thus circular, the circle is not necessarily vicious. Conceivably an iterative procedure, in which h_0 is estimated from a prior estimate of b (more precisely of $S(\dot{b})/b^2$) and the latter is then revised in light of $\hat{h}_0$, may bring off convergence to a more accurate estimate of b than the one with which the iteration begins. But success at this will clearly be a delicate business, critically dependent on both the sampling details of the particular parameters at issue and an astute choice of input to the iteration.

Equation 2 does not properly characterize any version of ridge regression that selects h in light of sample data, insomuch as h now has a sampling distribution that is not independent of $\dot{b}$. Instead, by some routine algebra applied to the analysis of $(\tilde{b} - b)^2$ noted prior to Equation 2, we find that

$$S(\tilde{b}) = S(\dot{b})\{(1 - h_0)^{-1} \times \text{Exp}[(h - h_0)^2] + h_0\} + C \quad (4)$$

where

$$C =_{\text{def}} \text{cov}(\dot{b}^2, h^2) + 2b\,\text{cov}[\dot{b}, h(h - 1)],$$

while h_0 is still defined by Equation 3 and cov designates sampling covariance. The quantity

$\operatorname{Exp}[(h - h_0)^2]$ is just the expected squared error $S(\hat{h}_0)$ of h construed as an estimate $\hat{h}_0$ of h_0, so Equation 4 can be rewritten as

$$S(\tilde{b})/S(\dot{b}) = h_0 + (1 - h_0)^{-1}S(\hat{h}_0) + C/S(\dot{b}). \quad (5)$$

Terms $S(\hat{h}_0)$ and $C/S(\dot{b})$ in Equation 5 can be analyzed further, but there is little present point in doing so except to claim without proof that all terms in Equation 5 are nonnegative unless ρ_{yx}^2 is less than N^{-1}, in which case it is possible for $C/S(\dot{b})$ to assume a small negative value. The present import of Equation 5 is simply that ridge regression can degrade estimational accuracy as well as enhance it. For regardless of how tiny h_0 may be, $S(\tilde{b})$ will be larger than $S(\dot{b})$ unless $S(\hat{h}_0)$ is sufficiently small. Since the sampling distribution of h $(=\hat{h}_0)$ under any specific method H for generating h from the sample data will be dependent on N and ρ_{yx}^2, one must anticipate that even if $S(\tilde{b})$ is less than $S(\dot{b})$ under some values of these parameters, the superiority order will be reversed in other regions of $\langle N, \rho_{yx}^2\rangle$. But if ridge-regression variant H does not dominate OLS over the entirety of parameter space—and it appears exceedingly unlikely that any choice of H can—then it is irresponsible to advocate and foolish to use H in preference to OLS unless we know what the regions of $\langle N, \rho_{yx}^2\rangle$ space are in which $S(\tilde{b})/S(\dot{b})$ is, respectively, appreciably less and appreciably greater than unity, as well as how large the difference from unity in each region tends to be. Only then will we be positioned to make rational judgments about which method, OLS or ridge-regression variant H, is most plausibly the more efficient for the particular application at hand. (In principle, choice of an estimator should be a complex process that involves prior credibilities and decision-theoretic utilities as well as sampling probabilities. But determining the relevant sampling distributions conditional on the relevant parameters is not merely an essential step in the decision process; it is the one step that we can actually execute in practice without flagrant appeal to vague and probably idiosyncratic intuitions.)

Insomuch as any choice of shrinkage parameter h $(=\hat{h}_0)$ may be taken via Equation 3 to define an estimate $\hat{\rho}_{yx}^2$ of correlation parameter ρ_{yx}^2 (i.e., one replaces h_0 by h, ρ_{yx}^2 by $\hat{\rho}_{yx}^2$, and solves for the latter), whereas conversely any estimate of ρ_{yx}^2 can be converted by Equation 3 into an estimate of h_0, each variant H of single-predictor ridge regression is grounded at least implicitly on some technique for estimating ρ_{yx}^2 without benefit of a prior ridge-regression estimate of b. (Even when ridge regression seeks to refine its solution by iteratively alternating between estimation of ρ_{yx}^2 and of b, there must always be a starting $\hat{\rho}_{yx}^2$.) It is not clear that there are many cogent ways to do this; in fact, there is probably only one that can be considered operationally preferable at this time, namely, the classic bias-corrected OLS estimate (Wherry formula)

$$\dot{\rho}_{yx}^2 =_{\text{def}} \begin{cases} 1 - (N-1)(N-2)^{-1}(1 - r_{yx}^2) & \text{if greater than 0,} \\ 0 & \text{otherwise,} \end{cases}$$

in which r_{yx} is the observed sample correlation.[4] [Note that when $\rho_{yx} \neq 0$, $\dot{\rho}_{yx}^2$ is always less than r_{yx}^2 and becomes zero when $r_{yx}^2 \leq (N-1)^{-1}$.] Let $\tilde{b}^*$ be the b estimate computed by the variant H^* of ridge regression that derives h via Equation 3 from ρ_{yx}^2 estimate $\dot{\rho}_{yx}^2$. I now submit as a working hypothesis that under almost any parameter point $\langle N, \rho_{yx}^2\rangle$, if $S(\tilde{b}^*) \geq S(\dot{b})$, then also $S(\tilde{b}) \geq S(\dot{b})$ for any variant of ridge regression other than H^*. That is, I suggest that the region of parameter space in which some variant of ridge regression is superior to OLS is roughly included in the region of superiority for ridge-regression variant H^*. (Arguments can be given to support this conjecture, but since they are inconclusive I forgo them here.) If so, study of $\tilde{b}^*$'s sampling behavior will reveal the conditions under which ridge regression can improve upon OLS.

Although it does not seem possible to derive an analytically exact value for $S(\tilde{b}^*)/S(\dot{b})$ given $\langle N, \rho_{yx}^2\rangle$, this quantity can be determined closely enough by Monte Carlo simulation. Table 1 reports such results for a rather broad spectrum of $\langle N, \rho_{yx}^2\rangle$ values,

[4] When x is only one of m orthogonal predictors, r_{yx} is the sample's partial correlation between x and y (see Footnote 3), and $N - 2$ in this formula for $\dot{\rho}_{yx}^2$ becomes $N - m - 1$.

Table 1
Monte Carlo Approximations to Ridge-Regression Inefficiency Ratio $S[\tilde{b}^*]/S[\dot{b}]$ *as a Function of Sample Parameters* N *and* ρ^2

$N \times \rho^2$	N = 10	20	50	100	500
0	.40 (.65)	.35 (.67)	.35 (.69)	.36 (.68)	.34 (.69)
.3	.59 (.59)	.54 (.62)	.52 (.60)	.55 (.59)	.52 (.60)
.6	.74 (.54)	.72 (.55)	.70 (.56)	.70 (.56)	.70 (.54)
.9	.88 (.46)	.85 (.46)	.83 (.50)	.87 (.52)	.86 (.51)
1.2	1.03 (.42)	.98 (.42)	.96 (.42)	.97 (.45)	.92 (.46)
1.5	1.16 (.33)	1.08 (.37)	1.04 (.38)	1.09 (.40)	1.07 (.40)
1.8	1.21 (.28)	1.23 (.33)	1.21 (.35)	1.11 (.34)	1.15 (.36)
2.1	1.33 (.26)	1.18 (.26)	1.19 (.30)	1.22 (.32)	1.21 (.35)
2.4	1.38 (.19)	1.26 (.24)	1.31 (.28)	1.19 (.28)	1.28 (.26)
2.7	1.44 (.18)	1.43 (.20)	1.37 (.25)	1.39 (.25)	1.36 (.27)
3.0	1.42 (.13)	1.40 (.21)	1.36 (.18)	1.41 (.22)	1.34 (.26)
4.0	1.47 (.07)	1.49 (.12)	1.44 (.15)	1.43 (.16)	1.53 (.16)
5.0	1.37 (.02)	1.55 (.06)	1.49 (.10)	1.54 (.10)	1.50 (.10)
6.0	1.28 (.00)	1.36 (.02)	1.42 (.05)	1.48 (.06)	1.52 (.07)
7.0	1.12 (—)	1.38 (.01)	1.42 (.03)	1.41 (.04)	1.48 (.05)
8.0	1.08 (—)	1.24 (.00)	1.40 (.02)	1.37 (.02)	1.43 (.04)
10.0		1.19 (—)	1.34 (.01)	1.37 (.01)	1.37 (.02)
12.0		1.10 (—)	1.24 (.00)	1.27 (.00)	1.36 (.01)
16.0		1.03 (—)	1.14 (—)	1.18 (—)	1.20 (.00)
20.0			1.11 (—)	1.13 (—)	1.17 (—)
30.0			1.04 (—)	1.08 (—)	1.10 (—)
40.0			1.02 (—)	1.04 (—)	1.06 (—)

Note. Approximate probabilities that the sample-computed estimate $\dot{\rho}^2$ of ρ^2 is zero are in parentheses. Each tabled entry is the mean of $(\tilde{b}_i^* - b_i)^2$ divided by the mean of $(\dot{b}_i - b_i)^2$, in a set of 1,000 samples (P_i) of size N and stipulated population parameter ρ^2. Each P_i was obtained by the following algorithm: (a) Generate N-termed vectors $\mathbf{x}_i$ and $\mathbf{e}_i$ by elementwise independent production from a random generator whose distribution is unit normal. (b) Compute the variance V_i of $\mathbf{x}_i$, determine $b_i = (b_i^2)^{\frac{1}{2}}$ by solving $\rho^2 = b_i^2 V_i/(b_i^2 V_i + 1)$, and derive $\mathbf{y}_i = b_i\mathbf{x}_i + \mathbf{e}_i$. (c) Compute ordinary least squares estimate $\dot{b}_i$ and ridge-regression estimate $\tilde{b}_i^*$ of b_i from sample vectors $\mathbf{x}_i$ and $\mathbf{y}_i$ by the appropriate formulas. Note that the sampling model for this simulation is random criterion residuals given a fixed predictor distribution.

together with the probability that $\dot{\rho}_{yx}^2 = 0$ at each parameter point. (Note that the table's rows are indexed by the product of N and ρ_{yx}^2 rather than by the latter alone). Estimator $\tilde{b}^*$ is more efficient than $\dot{b}$ so long as $\rho_{yx}^2 < N^{-1}$ and is substantially so when ρ_{yx}^2 is closer to zero than to N^{-1}. But $S(\tilde{b}^*)$ begins to exceed $S(\dot{b})$ when ρ_{yx}^2 becomes only fractionally greater than N^{-1}, reaching inefficiencies half again as large as those of OLS. Moreover, although $\tilde{b}^*$ returns essentially to parity with $\dot{b}$ as ρ_{yx}^2 becomes sufficiently large, the range of ρ_{yx}^2 values over which $\tilde{b}^*$ is appreciably inferior to $\dot{b}$ for any given N is much larger than its range of superiority.

Table 1 makes plain that $S(\tilde{b}^*)$ is less than $S(\dot{b})$ for just those combinations of N and ρ_{yx}^2 under which zero $\dot{\rho}_{yx}^2$ has rather high probability. Indeed, it is precisely this rounding of the bias-corrected r_{yx}^2 up to zero that makes $\tilde{b}^*$ superior to $\dot{b}$ in the $\rho_{yx}^2 < N^{-1}$ region. It is not so clear, however, that successful ridge regression depends on the zero bound on ρ_{yx}^2. Instead, what seems to be the fundamental nature of ridge regression, and more generally of any "shrinkage" adjustment of any sample-based estimator, can be insightfully idealized as follows: Given some function $\hat{\theta}$ of sample data construed to estimate a scalar population parameter θ, select for each sample size N some fixed bounded interval I_N of $\hat{\theta}$ values, define $\tilde{\theta}$ to be the sample function whose value respectively equals the midpoint θ_I of I_N or the value of $\hat{\theta}$ according to whether the latter is or is not in I_N, and consider the θ-estimation efficiency of $\tilde{\theta}$ versus $\hat{\theta}$ as a

function of θ with fixed N. It is not hard to see that so long as I_N is neither too wide nor too narrow and the sampling distributions of $\hat{\theta}$ have reasonably orthodox properties (notably, tailing off in both directions from a center in the vicinity of θ), $S(\bar{\theta})/S(\hat{\theta})$ has a minimum value less than unity at a value of θ near θ_I from which, as θ increases (or decreases), $S(\bar{\theta})/S(\hat{\theta})$ first increases to a maximum greater than unity and then subsides to an asymptotic value of unity. The details of this relative-inefficiency function are rather sensitive to the width of I_N; and one can seek a width that yields a useful θ region within which $S(\bar{\theta})/S(\hat{\theta})$ is appreciably less than unity while the degree to which $S(\bar{\theta})$ exceeds $S(\hat{\theta})$ elsewhere is not large enough to be a significant loss. How closely such an ideal choice of I_N can be attained in various particular cases I have no idea.

The simple model I have just described can obviously be generalized in various ways. For example, more than one shrinkage interval can be adopted simultaneously; the width and placement of I_N can be allowed to depend in part on sample properties additional to N; and shrinkage toward θ_I can be a graded function of the difference between $\hat{\theta}$ and θ_I rather than all or none according to whether $\hat{\theta}$ is in I_N. Even so, the all-or-none model makes clear the rationale of this approach. In particular, it removes the prima facie implausibility that the logic of ridge regression might apply with equal success to improving for example the OLS estimate of the population mean. For estimation of almost any statistic θ, what the generic shrinkage method can achieve by modifying a received θ-estimator $\hat{\theta}$ is a pocket of increased estimational accuracy (compared with the accuracy of $\hat{\theta}$) in the vicinity of one or more stipulated θ values θ_I at the price of decreased efficiency when θ's actual value lies elsewhere, the greatest losses occurring just outside of the region(s) of gain. Evidently, the occasions most appropriate for exploiting this technique are those in which our prior credibility distribution for θ is concentrated in the vicinity of θ_I.

Multicollinearity Revisited

If my analysis above is sound, it greatly clarifies ridge regression's potential value for the predictor multicollinearity problem. Returning to where the section entitled Preliminaries leaves off, let $f_1, \ldots, f_m$ be the (data space) "principal factors" of the X configuration, that is, f_i is the ith principal component of the predictor distribution rescaled to have unit variance. Then $X = F(\mathbf{D}_\lambda\mathbf{T})$ or $F = X(\mathbf{T}'\mathbf{D}_\lambda^{-1})$, in which rotation matrix $\mathbf{D}_\lambda\mathbf{T}$ is the product of an orthonormal matrix $\mathbf{T}$ with the diagonal matrix $\mathbf{D}_\lambda$ of roots of the X configuration's covariance matrix. It can be shown that $\lambda_m \leq \sigma_{x_i}^2(1 - R_i^2)$, in which R_i is the multiple correlation of any x_i in X with the other predictors and λ_m is the smallest predictor root; hence if any of the x_i are highly predictable from the rest, λ_m^{-1} will be very large. Moreover,

$$S(\hat{b}_{Xi}) = \sum_{j=1}^{m} t_{ij}^2\lambda_j^{-1}S(\hat{b}_{Fj}), \qquad (6)$$

where $\hat{b}_{Xi}$ ($\hat{b}_{Fi}$) is either the OLS or ridge-regression estimate of the ith coefficient in y's regression on X (F). Since

$$\sum_{i=1}^{m} t_{ij}^2 = \sum_{j=1}^{m} t_{ij}^2 = 1,$$

Equation 6 says that $S(\hat{b}_{Xi})$ is a weighted average of the $S(\hat{b}_{Fj})$ *after these have been inflated (or shrunk) by the associated* λ_j^{-1}; nor is there any tendency for the weights t_{ij}^2 to countervail that inflation, insomuch as the average t_{ij}^2 across $i = 1, \ldots, m$ is m^{-1}, the same for all j. The large uncertainty in the OLS estimate of b_{Xi} under predictor near-multicollinearity lies not in any special difficulty in estimating b_{Fj} when f_j's associated root is small—for $S(\hat{b}_{Fj})$ is the same for all j, namely, $N^{-1}\sigma_e^2$—but in the rotation from F to X, and is fully as troublesome for ridge regression as it is for OLS estimation.

Even so, there is a practical admonition for ridge regression in Equation 6. Let γ_{Fj} be the inefficiency ratio $\gamma_{Fj} =_{\text{def}} S(\bar{b}_{Fj})/S(\hat{b}_{Fj})$, and for simplicity assume—as will generally obtain closely enough when the predictors are nearly multicollinear—that for a given i, $t_{ij}^2\lambda_j^{-1}$ is much larger for those j in a subset J_i of indices $j = 1, \ldots, m$ than it is for the rest. Then $S(\bar{b}_{Xi})/S(\hat{b}_{Xi})$ is approximately equal to the average value of γ_{Fj} over just the j in J_i. (Although this is a considerable simplification

of the exact relationship, it fairly characterizes its essential nature, especially when the γ_{Fj} have a small coefficient of variation across J_i.) Given predictor near-multicollinearity, moreover, J_i will almost always comprise a subset of those j for which λ_j is very small. It follows that if ridge regression can substantially improve the estimation of regression coefficients for *some* of the predictor distribution's principal factors and can focus its prowess on those with the smallest associated roots, it may yet manage to provide succor specifically for the multicollinearity problem.

This hope is not entirely forlorn. Given fixed predictor scales, fixed regression coefficients, and fixed residual criterion variance, it is easily seen that the squared population correlation between y and any predictor x is an increasing function of σ_x^2 and, conversely, approaches zero as σ_x^2 approaches zero. Moreover, although the principal factors in Equation 6 have unit variances by scale stipulation, there is reason to suspect that very small predictor roots often symptomatize predictor configurations that, in an obscurely intuitive sense, allocate only diagnostically negligible variance to the dimensions of predictor space corresponding to these roots. It is impossible to be clear on this point without wallowing in a complicated, assumption-laden account of causal structure, scaling conventions, and the origins of predictor scores; but we can expect that some tendency exists—just how strong a one remains an important open question—for very small λ_j to be associated with especially small $\rho_{yf_j}{}^2$. And we have already seen that ridge regression's best chance to improve upon OLS is when $\rho_{yf_j}{}^2$ is *very* small. Unhappily, even when $\rho_{yf_j}{}^2$ is very small, if it is not small enough, $\tilde{b}_{Fj}$ will be *less* accurate than $\dot{b}_{Fj}$, and this loss will be amplified in $\tilde{b}_{Xi}$ by λ_j^{-1}.

If ridge regression is to be applied to highly intercorrelated predictors, one clear recommendation emerges. One should *not* routinely attenuate the OLS-estimated regression coefficients specifically for the predictors' principal factors (or principal components) with very small associated roots. Instead ridge regression should be applied only to those principal factors f_j whose (partial) squared correlation with the criterion can plausibly be inferred from the sample correlation and any other relevant information to be small enough for $S(\tilde{b}_{Fj}) < S(\dot{b}_{Fj})$. If the f_j that so qualify also correspond to very low λ_j, as seems not unlikely but still far from certain, then that same degree of increased efficiency—or loss thereof if we have misjudged—can be passed along to ridge-regression estimates of the b_{Xi}.

Conclusions

The import of my present argument is that we can expect to benefit from ridge regression—and otherwise to lose—only given special parametric circumstances whose presence in real-life applications cannot yet be reliably ascertained. Yet a number of Monte Carlo simulations in the ridge-regression literature (see especially Dempster, Schatzoff, & Wermuth, 1977, as well as earlier studies referenced by Price, 1977) have shown one or more variants of ridge regression to be impressively superior to OLS under the parameters tested. However, the design of these simulations—assigning population regression weights to multiple predictors also given rather high degrees of near-multicollinearity—does not permit easy comparison to Table 1; and there is no inconsistency between my present Monte Carlo conclusions and the published multivariate simulations if the manner in which these have been constructed produces parametric partial correlations between the criterion and the predictors' near-zero-root principal factors so small as to lie within the particular method's success region for the sample sizes tested. But if this is so—and it must be so if my preceding argument is not badly flawed—then the practical significance of these simulations remains severely problematic. They *may* exhibit a reliable tendency in all multivariate data arrays, real and artificial alike, for small-root principal factors to have vanishingly small correlations with an outside criterion. (If so, ridge regression can be operationally recommended just as soon as we determine how small a root is small enough.) But alternatively, they may well illustrate merely that artificial data arrays often approach ideal simplicities much more closely than do real data. Unless the *ways* in which multivariate distributions arise empirically are suitably reflected in simulation

studies, one must be exceedingly cautious in generalizing from one to the other.

I am prepared to argue that the only natural circumstances in which ridge regression can expect a reliable association between minuscule predictor roots and vanishing criterion correlations are roughly those in which the predictor variables are themselves causal sources of the criterion. In all other cases, the appropriate way to deal with intercorrelated predictors is through inferential factor analysis. But that is a story for another occasion. For now, the salient summary is that although ridge regression and other shrinkage techniques hold promise for a considerable diversity of specialized purposes, their employment also incurs a distinct risk of appreciable loss; and we still lack operationally effective criteria, or even a rudimentary theory thereof, for judging when the risk is worth taking. In particular, before extant variants of ridge regression are urged upon unwary statistical consumers, it is imperative that we acquire some knowledge about whatever relation may obtain in *empirical* data arrays between predictor roots and the criterion's correlations with the corresponding dimensions of predictor space.

For applied statistical estimation, ridge regression's day may come. But it has not come yet.

References

Dempster, A. P., Schatzoff, M., & Wermuth, N. A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, 1977, *72*, 77–91.

Mayer, L. S., & Wilke, T. A. On biased estimation in linear models. *Technometrics*, 1973, *15*, 497–508.

Price, B. Ridge regression: Application to nonexperimental data. *Psychological Bulletin*, 1977, *84*, 759–766.

Received October 3, 1977 ■

Call for Nominations

The APA Publications and Communications Board invites nominations for the editorship of *Psychological Bulletin* for the years 1981 through 1986. R. J. Herrnstein is the incumbent editor. The new editorial appointment will be made this year in order to provide continuity in publication of the journal. The editor-elect will start to receive manuscripts in 1980 to prepare for issues published in 1981, the first year of the new editorial term.

To nominate candidates, prepare a brief statement of one page or less in support of each nomination. Submit nominations no later than April 1, 1979, to Anita DeVivo, P&C Board Liaison, APA, 1200 Seventeenth Street, N.W., Washington, D.C. 20036.

Psychological Bulletin
1979, Vol. 86, No. 2, 250-259

Intellectual Functioning in Duchenne Muscular Dystrophy: A Review

Nicholas J. Karagan
Department of Pediatrics
University of Iowa

Duchenne muscular dystrophy has traditionally been thought to be a primary disease of muscle, but recently it has been suggested that it may be secondary to a neuronal defect or to a generalized disorder of protein synthesis and membrane. However, to date there is no proof to unequivocally support any of these theories. A higher incidence of mental retardation and decreased intellectual functioning has been reported in the medical literature for Duchenne muscular dystrophy patients than for normals or other control groups. Recently there has been strong evidence to suggest that verbal ability, as reflected by the Wechsler Intelligence Scale for Children Verbal scale IQ, may be more commonly and significantly impaired in Duchenne muscular dystrophy patients than is nonverbal ability, as reflected by the Performance scale IQ. This article presents a comprehensive review of intellectual functioning in Duchenne muscular dystropy. The intent is to provide a basis for future attempts to relate the intellectual deficit in Duchenne muscular dystrophy to neuropsychological and neurobiological parameters of the disease.

The muscular dystrophies are one class of inherited myopathies that are characterized by progressive muscular weakness and degeneration of muscle tissue without apparent cause in either the peripheral or central nervous system. In this regard they are distinguished from another class of myopathies, the spinal muscular atrophies, that involve progressive weakness and degeneration of muscle secondary to degeneration of anterior horn cells (Moosa, 1974). Duchenne muscular dystrophy has traditionally been considered a primary disease of muscle, but recently it has been suggested that it may be secondary to a neuronal defect (McComas, Sica, & Currie, 1970, 1971; Moosa, 1974) or to a generalized disorder of protein synthesis and membrane (Ionasescu, 1975). However, to date there is no proof to unequivocally support any of these theories.

Of the various types of muscular dystrophies, Duchenne or pseudohypertrophic muscular dystrophy is by far the most common, as well as the most rapidly progressive. Its incidence is 279 per 1,000,000 male births (Hanson & Zellweger, 1968). Death usually occurs from inanition or cardiopulmonary disease between 15 and 20 years of age. Usually, only males are affected by the disease, which is X-linked recessive. Thus, female carriers have the theoretical probability that 50% of their male offspring will be affected and that 50% of their female offspring will be carriers. However, about 30% of the affected males are the result of spontaneous mutation (Hanson & Zellweger, 1968). Duchenne muscular dystrophy does occur in rare instances in females, such as in Turner's syndrome (sex chromosomal complement XO) or in carriers in which the Lyon effect (Lyon, 1961, 1966; Murphy & Thompson, 1969) is thought to permit its clinical expression. Diagnosis of the disease is based

This study was supported in part by U.S. Public Health Service Grant 07-H-000044. The author wishes to thank John Sorensen, Lynn Richman, and Hans Zellweger for their valuable comments on earlier drafts of this article.

Requests for reprints should be sent to Nicholas J. Karagan, Department of Pediatrics, S252 Hospital School Building, University of Iowa, Iowa City, Iowa 52242.

		Performance	Stage
Ambulatory	Can climb stairs	Can climb stairs without aid of railing	I
		Can climb stairs with aid of railing with mild exertion	II
	Cannot climb stairs without another person's assistance	Can climb stairs with aid of railing but slowly, with labor, and cumbrously	III
		Can assume standing position from standard chair independently	IV
		Cannot assume standing position from standard chair independently	V
Predominantly in wheelchair but can walk 50 meters on level, usually with braces or other assistance			VI
Wheelchair existence	Independent transfer activity (from wheelchair to bed, toilet, etc.)		VII
	Dependent in transfer activities	Can maintain erect sitting posture independently	VIII
		(a) Cannot maintain erect sitting posture independently (b) Cannot raise arms 20 cm off arm rests	IX
Bed existence: Cannot use wheelchair			X

Figure 1. Functional classification of Duchenne muscular dystrophy. (From "Psychometric Studies in Muscular Dystrophy Type IIIa (Duchenne)" by H. Zellweger and J. W. Hanson *Developmental Medicine and Child Neurology*, 1967, 9, 576–581. Copyright 1967 by Spastics International Medical Publications. Reprinted by permission.)

on a positive family history, clinical symptoms, elevated serum creatine phosphokinase, electromyography, and ribosomal protein synthesis (determined from a muscle biopsy). Figure 1 summarizes the 10 stages of the disease, which are based on the ambulatory and transfer capabilities of the patient (Zellweger & Hanson, 1967). As can be seen, upper limb involvement progresses slowest, and finger dexterity is intact until the very latest stages, indicating that manual tasks can be accomplished without interference of the disease through Stage 5 and with minimal to moderate involvement through Stage 8.

A higher incidence of mental retardation and decreased intellectual functioning has been reported in the medical literature for Duchenne muscular dystrophy patients than for normals or other control groups. The reason for this finding is thought to be some type of abnormal central nervous system functioning. Recently, there has been strong evidence to suggest that verbal ability, as reflected by the Wechsler Intelligence Scale for Children (WISC) Verbal IQ, may be more commonly and significantly impaired than is nonverbal ability, as reflected by the WISC Performance IQ.

The purpose of the present article is twofold: first, to introduce a recent, comprehensive review of intellectual functioning in Duchenne muscular dystrophy into the psychological literature, and second, to highlight some of the methodological issues involved in the assessment of intelligence in this population. It is important to provide a basis for future attempts to relate intellectual deficit in Duchenne muscular dystrophy to neuropsychological and neurobiological parameters of the disease. I hope this approach will also prove useful in the study of brain–behavior relationships in other physical diseases that

Table 1
Summary of Studies Since 1960 of Intellectual Functioning in Duchenne Muscular Dystrophy

Authors	Year	*N*	Intelligence tests	Duchenne *M* IQ	IQ range	Control group *M* IQ	Age range in years
Sherwin & McCully	1961	15	WISC Verbal	103	90–120	None	10–14
			WISC Perf.	Not reported			
Allen & Rodgin	1960	30	S-B WISC WAIS IIS Rorschach	82	14–117	None	2–23
Worden & Vignos	1962	38	S-B	83	46–134	Unaffected sibs: 110 Diabetic patients: 107 Sibs of diabetics: 109 Spinal muscular atrophy: 118	4–15
Gamstrop & Smith	1964	10	S-B	91 (4 less than 82)	46–122	6 with denervation: 106 (1 less than 82)	3–17
Schorer	1964	28	S-B WISC WAIS Others	79	47–125	None	5–28
Dubowitz	1965	27	S-B	68	42–118	None	8–16
Zellweger & Niedermeyer	1965	45	S-B	83	42–131	None	3–16
		23 of 45	WISC Verbal	89	Not reported		Not reported
			WISC Perf.	97			
Zellweger & Hanson	1967	38	S-B WISC WAIS	84	56–127	None	2–18
Cohen, Molnar, & Taft	1968	211	Variety & estimated	86	Not reported	Unaffected siblings: 4 cases of mental retardation	Not reported
Nakao, Kito, Muro, Tomonaga, & Mozai	1968	86	WISC	Not reported; 13% borderline; 22% defective	Not reported	None	Less than 15

Table 1 (*continued*)

Authors	Year	*N*	Intelligence tests	Duchenne *M* IQ	IQ range	Control group *M* IQ	Age range in years
Prosser, Murphy, & Thompson	1969	52	S-B WISC WAIS	87	51–113	Unaffected sibs: 107 Unaffected brothers: 110	2–18
Rosman	1970	10	Not reported	Not reported; 6 of 10 had IQ of 27–75	27 (upper limit not reported)	None	2–18
Kozicka, Prot, & Wasilewski	1971	100	S-B WISC	76	70–110 (approx.)	Spinal muscular atrophy: 92	3–17
Michal	1972	15	S-B WISC	84	39–122	Postpolio boys: 97	6–14
Black	1973	25	WISC Verbal WISC Perf.	85 80	Not reported	None	Not reported
Marsh & Munsat	1974	16 (mildly impaired)	WISC Verbal WISC Perf.	85 98	65–111 67–128	Mild versus marked impairment	5–10
		18 (marked impairment)	WISC Verbal WISC Perf.	88 90	65–109 65–115		9–15
Florek & Karolak	1977	129	WISC S-B	79	30–127	Spinal muscular atrophy patients: 98	3–17
Karagan & Zellweger	1978	53	WISC Verbal WISC Perf.	81 88	58–108 51–118	$V–P \leq -7$ versus $V–P > -7$	5–9
		27 of the 53 with $V–P \leqslant -7$	WISC Verbal WISC Perf.	81 97	58–106 71–118		
		26 of the 53 with $V–P > -7$	WISC Verbal WISC Perf.	81 78	62–108 51–103		

Note. WISC refers to the Wechsler Intelligence Scale for Children; WAIS refers to the Wechsler Adult Intelligence Scale; S-B refers to the Stanford-Binet Intelligence Scale; IIS refers to the Infant Intelligence Scale. Perf. stands for the Performance scale of the WISC, and V–P stands for Verbal–Performance discrepancy on the WISC.

have behavioral correlates. Reviewed in detail are studies since 1960 of intellectual functioning in Duchenne muscular dystrophy.

Studies Prior to 1960

A higher prevalence of mental retardation than is found in the general population was documented by numerous reports prior to 1960 (Becker, 1953; Del Carlo Giannini & Marcheschi, 1959; Duchenne, 1868; Erb, 1891; Gowers, 1879; Zellweger, 1946) that studied only patients with Duchenne muscular dystrophy. Several earlier studies, however, did not report a higher incidence of mental retardation. They either reported on a mixed group of dystrophies (Bell, 1943; Morrow & Cohen, 1954; Schoelly & Fraser, 1955) or found the degree of intellectual impairment to be relatively mild (Walton & Nattrass, 1954). These earlier studies frequently involved clinical rather than psychometric evaluation of the retardation.

Studies Since 1960

Decline in General Intelligence

Table 1 contains a summary of studies since 1960 of intellectual functioning in Duchenne muscular dystrophy. There is only one, which reported on a group of Duchenne patients using formal psychometric assessment, that did not suggest a higher than normal incidence of mental retardation or a significantly lower than average overall IQ. Sherwin and McCully (1961) studied 15 children with Duchenne muscular dystrophy who ranged in age from 10 to 14 years. All were confined to a wheelchair. The verbal portion of the WISC was administered to all children. Mean Verbal IQ was 103 and ranged from 90 to 120.

All other studies have reported a lower than normal general or Full Scale mean IQ and/or a higher than normal incidence of mental retardation (Allen & Rodgin, 1960; Black, 1973; Cohen, Molnar, & Taft, 1968; Dubowitz, 1965; Florek & Karolak, 1977; Gamstrop & Smith, 1964; Karagan & Zellweger, 1978; Kozicka, Prot, & Wasilewski, 1971; Marsh & Munsat, 1974; Michal, 1972; Nakao, Kito, Muro, Tomonaga, & Mozai, 1968; Prosser, Murphy, & Thompson, 1969; Rosman, 1970; Worden & Vignos, 1962; Zellweger & Hanson, 1967; Zellweger & Niedermeyer, 1965). The mean IQs ranged from 68 to 91, with most about 1 *SD* below the general population average. Cohen et al. (1968), Florek and Karolak (1977) and Prosser et al. (1969) studied in detail the distribution of IQs in their samples. They concluded that the entire IQ distribution of the dystrophic children was shifted downward by 1 *SD* compared with the normal IQ distribution. Consequently, they reasoned that even those patients with an average or above-average IQ were reduced in intellectual ability compared with what would be expected if they did not have the disease.

Control Groups

The study of the intellectual functioning of relatives and affected and unaffected siblings has yielded some consistent and interesting results (Cohen et al., 1968; Kozicka et al., 1971; Prosser et al., 1969; Worden & Vignos, 1962). These studies uniformly reported higher mean IQs for the unaffected siblings and relatives (e.g., parents, aunts, and uncles) than for the affected siblings. Further, there was concordance in level of IQ between Duchenne patients and their affected siblings. The IQs of Duchenne patients also tended to follow the intellectual genetic history of the family, that is, severe retardation was found in patients from dull families, whereas normal intelligence was present in patients from bright families.

Groups of Duchenne patients have also been compared with a variety of control groups in the assessment of their level of intellectual functioning. Groups with chronic or physical incapacities such as diabetes mellitus (Worden & Vignos, 1962), spinal muscular atrophy (Florek & Karolak, 1977; Kozicka et al., 1971; Worden & Vignos, 1962), and postpoliomyelitis (Michal, 1972) all manifested average mean IQs, whereas Duchenne patients manifested below-average mean IQs.

These findings on relatives and other control groups are parsimonious, given the general finding that the distribution of IQ in

the Duchenne patients is shifted downward about 1 *SD* from the normal distribution.

Early Impairment of Verbal Intelligence

The conclusion gleaned thus far, that there is a generally decreased overall or Full Scale IQ in patients with Duchenne muscular dystrophy, is exceedingly impressive. Several investigators (Black, 1973; Prosser et al., 1969; Sherwin & McCully, 1961; Worden & Vignos, 1962; Zellweger & Hanson, 1967; Zellweger & Niedermeyer, 1965) have sought to associate a differential pattern of impairment with a lower Verbal than Performance IQ in patients with muscular dystrophy, but were either unsuccessful or did not explain the finding of a depressed Verbal IQ.

Marsh and Munsat (1974) were the first to document an early impairment of verbal intelligence in Duchenne muscular dystrophy. They studied 34 boys with Duchenne muscular dystrophy who ranged in age from 5 to 15 years. Of the subjects, 16 were mildly physically impaired, and 18 were moderately or severely disabled by the disease and required placement in a wheelchair. All were tested with the WISC. The mildly disabled group had a significantly lower Verbal scale IQ ($M = 85$) than Performance scale IQ ($M = 98$). However, for the moderately or severely disabled groups there was no significant difference between their Verbal scale IQ ($M = 88$) and their Performance scale IQ ($M = 90$). All children in the mild group were still ambulatory and, with one exception, were less than 10 years of age. Marsh and Munsat concluded that verbal intelligence was depressed and nonprogressive in these cases of Duchenne muscular dystrophy. On the other hand, Performance IQ tended to decrease with time as the physical disability interfered with the Performance scale items.

The authors noted that prior to their study there was no available published report of verbal and performance intelligence testing of dystrophic children using only the WISC. All previous studies had used two or more different intelligence tests to cover the age range in their samples. Further, in a detailed review, Marsh and Munsat pointed out that the various tests used were not truly equivalent (Barclay & Carolan, 1966; Hannon & Kicklighter, 1970) and that real differences may have been obscured by the use of several different measures.

Karagan and Zellweger (1978) attempted to replicate the findings of Marsh and Munsat and to study the pattern of Verbal and Performance IQ in this population in more detail. They studied intellectual functioning in a group of 53 boys with Duchenne muscular dystrophy. All the children were under 10 years of age and were ambulatory, that is, they were Stage 5 (Zellweger & Hanson, 1967) or less of their disease. All were administered the Verbal and Performance scales of the WISC; mean Verbal IQ (81) was significantly lower ($t = 4.48$, $p < .001$) than Performance IQ (88). Further, the range of Verbal IQs (58–108) was more constricted and the distribution more skewed than was the range (51–118) and distribution of Performance IQs. The mean Verbal–Performance discrepancy score was -7, significantly lower than the score for the standardization sample of the WISC, which was 0 (Seashore, 1951).

Two Patterns of Intelligence Test Performance

In an effort to analyze this pattern in more detail, Karagan and Zellweger (1978) dichotomized this sample first at the mean (-7) of the Verbal–Performance discrepancy distribution. Two distinct patterns of intelligence test performance emerged. Both groups demonstrated depressed mean Verbal IQs of 81. However, the first group, with a Verbal–Performance discrepancy score that was less than or equal to -7, had a mean Performance IQ of 97. The other group, with a Verbal–Performance discrepancy score that was greater than -7, also had a significantly lower than normal Performance IQ (78).

Relationship of IQ to Psychosocial Factors

Emotional Factors

Studies prior to 1960 (e.g., Morrow & Cohen, 1954) generally attributed any identi-

fied intellectual deficit in Duchenne muscular dystrophy to emotional factors associated with chronic and fatal illness. Although Allen and Rodgin (1960) and Florek and Karolak (1977) identified in the patients in their studies emotional components that could have had a depressing effect on their level of intellectual functioning, these components could not account for the overall depression of IQ, particularly the lower IQs.

Karagan and Zellweger (1978), Kozicka et al. (1971), Marsh and Munsat (1974), and Prosser et al. (1969) did not find the generally lower IQs of dystrophic patients to be the result of secondary or environmental effects of the disease. Kozicka et al. found an average IQ (92) in their control group of patients with spinal muscular atrophy who were about as equally chronically ill and physically disabled as the Duchenne patients, who had a mean IQ of 76. Karagan and Zellweger, Marsh and Munsat, and Prosser et al. concluded that psychosocial factors have little influence on the depressed IQs of Duchenne patients for several reasons. First, the depressed intellectual functioning is present at an early age. Second, most young children with Duchenne muscular dystrophy are no more than minimally restricted in mobility and thus are able to adequately explore their environment. And finally, most of the younger children in these studies either were enrolled in a regular school program or were placed in special programs solely on the basis of their lower mental functioning, and they did not manifest educational handicaps directly attributable to their physical disability.

Socioeconomic Status

No relationship has been demonstrated between socioeconomic status and IQ, except that between extreme classes the IQ differences are not unlike those for the general population (Cohen et al., 1968; Florek & Karolak, 1977; Prosser et al., 1969; Worden & Vignos, 1962; Zellweger & Niedermeyer, 1965). Prosser et al. commented that at both the low and middle socioeconomic levels, the dystrophic patients have lower IQs than do their normal sibs.

Relationship of IQ to Physical Factors

Serum Enzymes

Worden and Vignos (1962) found no relationship between IQ and creatinuria or serum aldolase. Prosser et al. (1969) looked at the relation between creatine phosphokinase (CPK) and IQ in relatives of Duchenne patients. In each family there was no difference in mean IQ between carriers and normals (both male and female). There was also a low nonsignificant correlation between the highest recorded CPK level of each carrier and her IQ. No known study has reported on the relationship between CPK level and IQ in the Duchenne patients themselves.

Sporadic Versus Familial Cases

Although Dubowitz (1965) found that a positive family history for the disease was much less frequent in patients of normal intelligence than in retarded patients, these findings were not supported by Prosser et al. or Zellweger and Niedermeyer (1965).

Severity of Disability

In several studies, IQ was unrelated to severity of disability (Allen & Rodgin, 1960; Black, 1973; Cohen et al., 1968; Kozicka et al., 1971; Nakao et al., 1968; Prosser et al., 1969; Worden & Vignos, 1962; Zellweger & Hanson, 1967), despite two reports suggesting a deterioration in intellectual functioning with disease progression (Dubowitz, 1965; Florek & Karolak, 1977). However, as noted previously, Marsh and Munsat (1974) reported a lower WISC Performance IQ in their patients who were physically more significantly impaired and concluded that the lower Performance IQ was a function of the interference of the physical disability with Performance scale items. This, of course, would also tend to produce a lower Full Scale IQ.

The findings of two reports suggest an interesting phenomenon with respect to Verbal IQ over time. Prosser et al. (1969) found a significant correlation of .38 between age and Verbal IQ in 39 of their patients. Karagan and Zellweger (1976; Karagan & Zellweger,

Note 1) reported on a preliminary test–retest study of the intellectual functioning of 22 children with Duchenne muscular dystrophy. The children were tested with the WISC first when they were under age 10 and at less than Stage 6 of their disease (ambulatory) and again when they were not ambulatory, an average of 48 months later. Again, consistent with their findings on 53 young dystrophic patients (Karagan & Zellweger, 1978), two groups emerged. On initial testing, the mean Verbal–Performance discrepancy score for the 22 patients was −11. The group of 11 patients with a Verbal–Performance discrepancy score that was less than or equal to −11 had an initial Verbal IQ of 79, but on retesting they had a mean Verbal IQ of 87 ($p < .10$). The mean Performance IQ of 99 on initial testing dropped to 88 ($p < .05$). The other group of 11 patients, which had on initial testing a Verbal–Performance discrepancy score that was greater than −11, had identical scores on test and retest—mean Verbal IQs of 79 and mean Performance IQs of 78. Neither group was significantly different in terms of age or stage of the disease at the time of the first or the second testing.

Abnormal Brain Functioning

Although most authors have concluded that the intellectual deficit is related to abnormal brain functioning (Allen & Rodgin, 1960; Cohen et al., 1968; Dubowitz, 1965; Karagan & Zellweger, 1978; Marsh & Munsat, 1974; Michal, 1972; Prosser et al., 1969; Rosman, 1970; Schorer, 1964; Worden & Vignos, 1962; Zellweger & Hanson, 1967), several studies have reported a high percentage of abnormal electroencephalogram (EEG) findings in patients with Duchenne muscular dystrophy (Florek & Karolak, 1977; Kozicka et al., 1971; Nakao et al., 1968; Zellweger & Niedermeyer, 1965). Further, these studies also reported a high correlation between the severity of the intellectual deficit and the severity of the EEG findings. This finding provides some support for the notion that the intellectual deficit is based on some physical parameter of the disease that affects central nervous system functioning as opposed to the notion that the intellectual deficit is solely secondary to psychosocial factors.

Progression of the Disease

Three studies provide some interesting but tentative and speculative data regarding the relation between IQ and progression of the disease. Rosman (1970) reported on 10 patients with Duchenne muscular dystrophy and found that both the clinical myopathy and the severity of histopathological changes paralleled intellectual functioning: The patients with the most severe involvement of muscle were those patients with the most severe degree of intellectual impairment. This relationship was independent of age of onset or of the duration or severity of the myopathy.

Nakao et al. (1968) found in 77 Duchenne patients that the rate of inferior intelligence was high (45%) when the disease was of 6 to 11 years duration and was low (20%) when the disease was of more than 11 years duration. They concluded that cases of inferior intelligence pursue an unfavorable course.

Karagan and Zellweger (1976; Karagan & Zellweger, Note 1) in their preliminary test–retest study of 22 Duchenne patients, found that the 11 patients with a Verbal–Performance discrepancy score of less than or equal to −11 progressed from stages 2 and 3 of the disease to Stages 8 and 9 in an average of 66 months. On the other hand, the 11 patients with a Verbal–Performance discrepancy of greater than −11 (i.e., both Verbal IQ and Performance IQ were depressed) progressed more rapidly, in an average of 38 months ($p < .10$).

Conclusion

Patients with Duchenne muscular dystrophy, traditionally thought to be a primary disease of muscle, manifest a higher incidence of mental retardation and decreased intellectual functioning than the general population. There is strong evidence that this decrease is the result of abnormal central nervous system functioning. Although it has recently been suggested that the disease may be secondary to a neuronal defect or to a generalized disorder of protein synthesis and membrane, no

proof of any of the proposed etiologies exists. The specific mechanisms underlying the disease itself and the accompanying intellectual deficits are unknown at the present time. However, a relationship between the intellectual deficit and the disease is an intriguing possibility.

The entire IQ distribution of patients with Duchenne muscular dystrophy appears to be shifted downward by about 1 *SD* compared with the normal IQ distribution. The IQs of these patients follow the genetic intellectual characteristics of the family, and there is high concordance between the intellectual levels of affected siblings. Duchenne patients typically have significantly lower IQs when compared with a variety of control groups who have chronic disease or physical incapacity. Verbal ability, as defined by the WISC Verbal IQ, is depressed at an early age and appears universally affected in Duchenne patients. Some patients, however, manifest a significant deficit in both Verbal and Performance IQ on the WISC.

There is no evidence to date that relates IQ deficit to serum enzymes, sporadic or familial etiology, age of onset of clinical symptoms, or the severity of disability. However, Performance scale IQ does decline over time as a function of the interference of physical disability with performance scale items. There is no evidence of any intellectual or cerebral deterioration. In fact, there is some suggestion that in some patients Verbal IQ manifests a modest increase over time.

Finally, there are several reports that suggest that the more severely affected a child's intellectual functioning is, the more rapidly the disease progresses. The evidence for this is by no means unequivocal yet.

What appears necessary at this juncture is to identify other neuropsychological characteristics of this disorder and, in turn, to relate these as well as the intellectual deficits to physical parameters of the disease. These factors, of course, do not appear to relate in a simple fashion. The study of neuropsychological aspects of other myopathic disorders is likewise important, particularly because these intriguing yet unfortunate human maladies have received relatively little attention in the psychological literature.

Reference Note

1. Karagan, N. J., & Zellweger, H. U. *IQ studies in Duchenne muscular dystrophy II: Test retest performance.* Paper presented at the meeting of the American Academy for Cerebral Palsy, New Orleans, September 1975.

References

Allen, J. E., & Rodgin, D. W. Mental retardation in association with progressive muscular dystrophy. *American Journal of Diseases of Children,* 1960, *100,* 208–211.

Barclay, A., & Carolan, P. A comparative study of the Wechsler Intelligence Scale for Children and the Standford-Binet Intelligence Scale, Form L-M. *Journal of Consulting Psychology,* 1966, *30,* 563.

Becker, P. E. *Dystrophia musculorum progressiva.* Stuttgart, West Germany: G. Thieme, 1953.

Bell, J. Nervous diseases and muscular dytrophies. In R. A. Fisher & L. S. Penrose (Eds.), *The treasury of human inheritance* (Vol. 4). London: Cambridge University Press, 1943.

Black, F. W. Intellectual ability as related to age and stage of disease in muscular dystrophy: A brief note. *Journal of Psychology,* 1973, *84,* 333–334.

Cohen, H. J., Molnar, G. E., & Taft, L. T. The genetic relationship of progressive muscular dystrophy (Duchenne type) and mental retardation. *Developmental Medicine and Child Neurology,* 1968, *10,* 754–765.

Del Carlo Giannini, G., & Marcheschi, M. Sui disturbi psyichici nella distrofia muscolare primitiva. *Sistema Nervoso,* 1959, *6,* 461–480.

Dubowitz, V. Intellectual impairment in muscular dystrophy. *Archives of Diseases of Children,* 1965, *40,* 296–301.

Duchenne, G. B. Recherches sur la paralysie musculaire pseudohypertrophique ou paralysie myosclerosique. *Archives Generales de Medecine,* 1868, *11,* 5–25.

Erb, W. Dystrophia muscularis progressive: Klinische und pathologischanatomische studien. *Deutsch Zeitschrift für Nervenheilkunde,* 1891, *1,* 173.

Florek, M., & Karolak, S. Intelligence level of patients with Duchenne type of progressive muscular dystrophy (PMD-D). *European Journal of Pediatrics,* 1977, *126,* 275–282.

Gamstrop, I., & Smith, M. EEG-fynd och testresultat vid myopati och denervation satrofi i barnaaldern. *Nordisk Medicin,* 1964, *72,* 998–1000.

Gowers, W. R. Clinical lectures on pseudo-hypertrophic paralysis. *Lancet,* 1879, *2,* pp. 1–2; 37–39; 73–75; 113–116.

Hannon, J. E., & Kicklighter, R. WAIS versus WISC in adolescents. *Journal of Consulting and Clinical Psychology,* 1970, *35,* 179–182.

Hanson, J. W., & Zellweger, H. U. The muscular dystrophies in Iowa. *Journal of the Iowa Medical Society,* 1968, *58,* 251–260.

Ionasescu, V. Distinction between Duchenne and

other muscular dystrophies by ribosomal protein synthesis. *Journal of Medical Genetics,* 1975, *12,* 40–52.

Karagan, N. J., & Zellweger, H. U. IQ studies in Duchenne muscular dystrophy II: Test retest performance. *Developmental Medicine and Child Neurology,* 1976, *18,* 251. (Abstract)

Karagan, N. J., & Zellweger, H. U. Early verbal disability in Duchenne muscular dystrophy. *Developmental Medicine and Child Neurology,* 1978, *20,* 435–441.

Kozicka, A., Prot, J., & Wasilewski, R. Mental retardation in patients with Duchenne progressive muscular dystrophy. *Journal of the Neurological Sciences,* 1971, *14,* 209–213.

Lyon, M. F. Gene action in the X-chromosome of the mouse. *Nature,* 1961, *190,* 372–373.

Lyon, M. F. Sex chromatin and gene action in the X-chromosome of mammals. In K. L. Moore (Ed.), *The sex chromatin.* Philadelphia, Pa.: Saunders, 1966.

Marsh, G. G., & Munsat, T. L. Evidence of early impairment of verbal intelligence in Duchenne muscular dystrophy. *Archives of Diseases of Children,* 1974, *49,* 118–122.

McComas, A. J., Sica, R. E. P., & Currie, S. Muscular dystrophy: Evidence for a neural factor. *Nature,* 1970, *266,* 1263–1264.

McComas, A. J., Sica, R. E. P., & Currie, S. An electrophysiological study of Duchenne dystrophy. *Journal of Neurology, Neurosurgery, and Psychiatry,* 1971, *34,* 461–468.

Michal, V. [The psychology of the child with progressive muscular dystrophy.] *Ceskoslovenska Psychiatrie,* 1972, *64,* 226–230. (*Psychological Abstracts,* 1972, *51,* No. 3547.)

Moosa, A. Muscular dystrophy in childhood. *Developmental Medicine and Child Neurology,* 1974, *16,* 97–111.

Morrow, R. S., & Cohen, J. Psychosocial factors in muscular dystrophy. *Journal of Child Psychiatry,* 1954, *3,* 70–80.

Murphy, E. G., & Thompson, M. W. Manifestations of Duchenne muscular dystrophy in carriers. In A. Barbeau & J. R. Brunette (Eds.), *Progress in neuro-genetics.* Amsterdam: Exerpta Medica Foundation, 1969.

Nakao, K., Kito, S., Muro, T., Tomonaga, M., & Mozai, T. Nervous system involvement in progressive muscular dystrophy. *Proceedings of the Australian Association of Neurologists,* 1968, *5,* 557–564.

Prosser, E. J., Murphy, E. G., & Thompson, M. W. Intelligence and the gene for Duchenne muscular dystrophy. *Archives of Diseases of Children,* 1969, *44,* 221–230.

Rosman, N. P. The cerebral defect and myopathy in Duchenne muscular dystrophy. *Neurology,* 1970, *20,* 329–335.

Schoelly, M. L., & Fraser, A. W. Emotional reactions in muscular dystrophy. *American Journal of Physical Medicine,* 1955, *34,* 119–123.

Schorer, C. E. Muscular dystrophy and the mind. *Psychosomatic Medicine,* 1964, *26,* 5–13.

Seashore, H. G. Differences between Verbal and Performance IQs on the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology,* 1951, *15,* 62–67.

Sherwin, A. C., & McCully, R. S. Reactions observed in boys of various ages (ten to fourteen) to a crippling, progressive, and fatal illness (muscular dystrophy). *Journal of Chronic Diseases,* 1961, *13,* 59–68.

Walton, J. N., & Nattrass, F. J. On the classification, natural history and treatment of the myopathies. *Brain,* 1954, *77,* 169–231.

Worden, D. K., & Vignos, P. J. Intellectual functioning in childhood progressive muscular dystrophy. *Pediatrics,* 1962, *29,* 968–977.

Zellweger, H. Über knochenveränderungen bei der dystrophia musculorum progressiva. *Annales Paediatrici (Basel),* 1946, *167,* 287–292.

Zellweger, H. U., & Hanson, J. Psychometric studies in muscular dystrophy type IIIa (Duchenne). *Developmental Medicine and Child Neurology,* 1967, *9,* 576–581.

Zellweger, H., & Niedermeyer, E. Central nervous system manifestations in childhood muscular dystrophy (CMD): I. Psychometric and electroencephalographic findings. *Annales Paediatrici,* 1965, *205,* 25–42.

Received October 11, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 260-275

Research on the Effects of Disconfirmed Client Role Expectations in Psychotherapy: A Critical Review

Paul Duckro
Malcolm Bliss Mental Health Center, St. Louis, Missouri

Don Beal and Clay George
Texas Tech University

This article critically examines the pervasive assumption found in psychotherapy literature that disconfirmation of client role expectations has been demonstrated to be a negative influence in psychotherapy. When the empirical literature is examined, this hypothesis does not appear to be as conclusive as has been suggested. In fact, the empirical studies are evenly divided in supporting this hypothesis. Implications for future research are discussed.

The expectations of client and therapist have for some time been postulated to be important influences in psychotherapy. In a monograph-length review and analysis of the early expectation research, A. Goldstein (1962b) extracted two types of expectations relevant to the study of psychotherapy. These two types were characterized as (a) prognostic expectations and (b) participant role expectations. The former were defined as the assessments of therapist and client regarding the probability of success in the therapeutic intervention. The latter were defined as the anticipations held by therapist and client regarding the behavior that will be displayed in the psychotherapeutic relationship by both participants. Researchers have hypothesized, for example, that clients bring to psychotherapy certain preconceived ideas about what the psychotherapist will do and how the client should behave.

The present review focuses on this second category—participant role expectations. More specifically, the review examines extant research that bears on the hypothesis that failure to confirm client expectations of the therapist's role results in negative consequences. In many ways, this hypothesis has already come to be accepted as fact by a large proportion of the psychological community. In the common wisdom, it enjoys the status of a virtually unquestioned assumption. Many such allusions to the "demonstrated" relationship of disconfirmed client role expectations and various dependent variables come across the desk of even the casual reader of research in psychotherapy. To demonstrate this point, three such allusions are presented here. That all three examples are drawn from recent, scholarly reviews of related areas of research only serves to emphasize the pervasiveness of the assumption that disconfirmation of client role expectations has been demonstrated to be a negative influence in psychotherapy.

First, Lorion (1974a) reported that "the importance of patient expectations to treatment variables . . . has been demonstrated" (p. 347). He felt sure enough of this statement to cite in support only seven representative studies that used five different types of dependent variables. Second, Baekeland and Lundwall (1975), in an extensive review of the psychotherapy dropout literature, wrote that "it is known that discrepant expectations about treatment promote drop-

---

Don Beal is now at Miami University.
Requests for reprints should be sent to Paul Duckro, Psychological Services, Malcolm Bliss Mental Health Center, 1420 Grattan Street, St. Louis, Missouri 63104.

ping out" (p. 758). This inference was supported by the citation of six studies, five of which were published prior to 1966. Thus, these authors showed little hesitation about accepting the validity of the disconfirmed expectations – negative consequences hypothesis. Third, Heitler (1976) made reference to the "substantial body of theory and research evidence" (p. 340) that indicates that some mutuality of patient–therapist role expectations is crucial. He was so confident of this generalization that he cited only six references in support.

Each of the preceding examples demonstrates the extent to which the hypothesized relationship of disconfirmed expectations and negative consequences in psychotherapy has been incorporated into the belief system of the field. The articles cited were not selected because they are unique. On the contrary, they represent a pervasive assumption among clinical researchers. The examples are particularly forceful because they appear in the context of high quality scholarly papers.

The present review was undertaken to examine more systematically whether the assumption of *proven* for the role expectation hypothesis is warranted by the available empirical evidence. The literature is discussed in the following order. First, descriptive expositions of client role expectations are presented briefly. Second, the theoretical and experimental background leading up to A. Goldstein's (1962b) keystone monograph is summarized. Third, the experimental literature since 1962 that deals with the question of the negative consequences of disconfirmed role expectations in psychotherapy is reviewed. Fourth, this literature is discussed in terms of its implications and ambiguities.

Descriptive Studies

A wide range of studies have empirically demonstrated the existence of client expectations about the therapist's role. Apfelbaum's (1958) work is perhaps the classic descriptive study in this area. Using outpatients of a university psychiatric clinic, Apfelbaum cluster analyzed the clients' Q sorts that reflected their pretherapy expectations of the therapist's attitudes and interview behavior. The resulting clusters suggested three major types of client role expectation: (a) the nurturant therapist—giving, protecting, and guiding without pushing or criticizing; (b) the model—well adjusted, diplomatic, a permissive listener but not protective; and (c) the critic—analytical, critical, and demanding considerable responsibility from the client.

Heine and Trosman (1960) identified two types of role expectations held by clients regarding the psychotherapy process. They called these two types of expectations the guidance model and the collaboration model. By far, most of the outpatients in their sample held expectations of psychotherapy that clearly fit the guidance model. They saw the therapist as the source of diagnostic information, of advice, and of medicine. Therapists, on the other hand, anticipated a therapy relationship closer to the collaborative model. They were generally less directive in style and were not oriented toward the use of diagnosis or psychopharmacological agents.

Finally, Begley and Lieberman (1970) broke down client role expectations in still another way. Using a modified version of a questionnaire developed by McNair and Lorr (1964), they identified two clusters of patient expectations of the therapist's role. One patient group anticipated an active, directive, but warm therapist. The other group expected a more passive, detached, objective therapist.

Using these and other approaches to the measurement of client expectations, a number of investigators have limited themselves to reporting the types of expectation predominant in their sample in the hope of determining the characteristic expectations of various populations. Many such reports indicated that the samples anticipated directive therapists. Patients expected the therapist to be warm but still firmly in control of the therapy session. For example, Thomas, Polansky, and Kounin (1955) found that university students characterized the helpful therapist as more directive, that is, willing to offer advice and to structure the situation. Chance (1957) studied a group of mothers who were in therapy concurrently with their children. She found that the

mothers projected their therapists' characteristics in idealized terms. They expected the therapist to function in the interviews as an active, directive, and supportive teacher. Cundick (1963) reported that college student outpatients expected an active and personally involved therapist. However, both clients and therapists agreed that the client had primary responsibility for the direction of the interview. This aspect of their expectation is different from the findings of the previous two studies. Finally, Tinsley and Harris (1976) found that their sample of college students did expect the therapist to be both expert and genuine in personal communication. Interestingly, this group did not generally expect that the therapy would be successful.

Correlates of Client Expectations

Not too far removed from the purely descriptive study of client expectations of therapist role, other researchers have used correlational designs to study the relationship of expectations to various client or psychotherapy variables.

Apfelbaum (1958) found the three role expectation factors in his study to be differentially associated with clients' Minnesota Multiphasic Personality Inventory profiles, rate of dropping out of psychotherapy, hours in therapy, and sex differences. Jacobs, Muller, Anderson, and Skinner (1972) studied the efficiency of several predictors of improvement in hospitalized patients. They found that expectations of a low-directive, less concerned, and less sensitive therapist were the strongest single predictor of negative outcome ($r = .31$).

Caine, Wijesinghe, and Wood (1973) compared persons with expectations of a directive, authoritative therapist to persons who expected the client to be a more collaborative partner in the process of therapy. They found that persons who expected the more directive therapist were significantly more externally directed, tended to be convergent rather than divergent thinkers, and scored higher on the Conservatism scale. Baldwin (1974) reported that university students who expected the therapist to be planned rather than spontaneous in his or her intrasession behavior were more likely to be repressors than sensitizers. In addition, repressors considered therapist personality to be a less important variable in therapeutic outcome than did sensitizers.

Based on his own work and a review of other research, Lorion (1974a, 1974b) concluded that expectations for psychotherapy are not related to socioeconomic status (SES). He reported that subjects from all SES levels tended to verbalize similar expectations of therapist behavior. Low SES clients did not express significantly stronger anticipations of an active, problem-solving therapist style. Supporting this position is Bent, Putnam, Kiesler, and Nowicki's (1975) report that their sample of relatively well-educated community mental health center outpatients expected to receive advice and medicine to solve their problems. These anticipations were not particularly different from those reported for a group of low SES outpatients by Overall and Aronson (1963). Garfield and Wolpin (1963) studied a group of young, predominantly low SES clients in which 67% expected types of therapist behavior that put considerable responsibility on the client to help himself or herself. However, Garfield and Wolpin also asked these clients what type of therapist behavior they preferred. In contrast to their expectation, the clients generally preferred to be given advice. This finding raised the question of the interaction between expectation and preference. The distinction between these two very different concepts has very seldom been made by researchers in the area. As we point out later in the article, the failure to make this distinction has been a source of ambiguity in the literature.

Perhaps the most justified conclusion from this series of descriptive and correlational studies is that expectations of therapist behavior will vary considerably among different samples and even within any given sample. In addition, the Overall and Aronson (1963) study hints that disconfirmed expectations might lead to negative results. Bent et al. (1975), in their descriptive re-

port, stated that this question—about the role of expectation in the outcome of treatment is the next logical question to be addressed by their research. They are correct, of course, in assigning a high priority to the question. The existence of client expectations per se is of little importance if the failure to acknowledge or confirm these expectations does not affect the therapy outcome or process. However, it must be recognized that a substantial body of research has been addressed to this question. In fact, as suggested earlier, the hypothesized relationship between disconfirmed expectations for therapist behavior and negative consequences in therapy has apparently been awarded factual status in much of professional psychology. The examination of the extant research that bears on this hypothesis is the crux of the present effort to ascertain whether its status as *demonstrated* is justified.

Theoretical and Experimental Background: Research Before 1962

Interest in the impact of client role expectations in psychotherapy was already evident by the early 1950s. Initial efforts provided at least explicit speculation on the importance of role expectations in psychotherapeutic treatment (e.g., Kamm & Wrenn, 1950; Seeman, 1949).

Kelly (1955) considerably elaborated the theoretical underpinnings of the early propositions regarding client role expectations in psychotherapy. He postulated that almost any client already holds a highly personalized conceptualization of the nature of the psychotherapy relationship and of the psychotherapist's role within that relationship even prior to the initiation of treatment. Kelly argued, on the basis of his theoretical position, that the psychotherapist must accept the client's preconception of the therapist's role, at least in the beginning stage of therapy. Failure to confirm the client's expectations results in confusion or disappointment on the part of the client. In Kelly's formulation, therefore, the therapist cannot ignore or reject without negative effect the client's anticipations, even though the therapist may attempt in the long run to change the role he or she initially accepts.

Danskin (1955) also posited that the effective therapist typically attempts to play the role expected of him or her by the client. On the other hand, Patterson (1958) argued that there are more and less effective ways to conduct therapy. He favored the low-directive, client-centered approach. Patterson's position was that therapists should not attempt to meet client expectations for therapist behavior. He admonished them to help clients, if necessary, learn to respond favorably to the low-directive therapy style.

At the same time, early empirical efforts to examine the expectation hypothesis were also under way. McGowan (1955) reported that clients in his sample experienced equivalent levels of satisfaction regardless of the style used by the counselors. Satisfaction was related, however, to the perceived expertise of the therapist. A major factor in the perception of a counselor as expert appeared to be his or her ability to form a close, facilitative relationship with the client. Frank, Gliedman, Imber, Nash, and Stone (1957) found that individuals assigned to the relatively unfamiliar (at that time) group method of therapy dropped out of treatment at a much higher rate than did those assigned to the more frequently anticipated individual therapy. The authors speculated that the group therapy experience was too sharply incongruent with the clients' expectations of psychotherapy to permit beneficial participation.

Biddle (1958) built his reasoning on the foundation of earlier work in social psychology. This literature suggested that if one person conforms to the role expectations held by a second person, the first will enjoy greater influence over the second than will a person who does not conform to those role expectations. Biddle hypothesized that analogue subjects would express low satisfaction with a taped therapist who failed to meet their expectations. The results supported the hypothesis only under one of two special conditions. First, when the therapist had some power to punish the client, nonconformity to client expectations led to less

satisfaction. Second, when the client was led to believe that the therapist expected to behave in much the same way as the client expected him or her to behave, then nonconformity by the therapist resulted in the expression of less satisfaction by the client.

Heine and Trosman (1960) used a sample that held predominant expectations of an active, directive doctor and a passively cooperative patient. Their therapists' expectations were generally much more along the lines of the collaborative model. The authors found that patients with the strongest expectations of a directive therapist were much more likely to drop out of therapy than were persons with somewhat less directive expectations. Although therapist styles were not closely controlled in this study, the results were suggestive. Similar findings were reported at about the same time by Skinner and Anderson (1959) and by Hankoff, Englehardt, and Freeman (1960).

Emboldened by these promising data, researchers began major studies, and theoretical propositions were offered. Chance (1959) reviewed the literature up to that point and reported the findings of her own extensive study. She concluded that "this mutuality of expectation may be one of the prerequisites to therapy" (p. 105). Wallach and Strupp (1960) hypothesized, on the basis of the available evidence, that if the client's expectations regarding the therapist were confirmed, then the therapy situation would appear more rewarding and the therapist would be perceived more positively. Therefore, they suggested that congruence between the type of assistance expected and the type of assistance proffered might be an important variable in psychotherapy.

Further, Lennard and Bernstein (1960) reported on their major investigation of the nature and interrelationship of role expectations and therapist–client communication in the psychotherapy interview. Their results suggested a significant relationship between the degree of dissimilarity of the participants' role expectations and the degree of dysfunction in the communication system. They concluded that when the expectations were very dissimilar, the resultant strain in the dyadic interpersonal system placed that system in proximate danger of disintegration. Finally, A. Goldstein (1962a, 1962b) comprehensively reviewed the literature that directly and indirectly bore on the influence of role expectations in psychotherapy. On the basis of those data, he adopted the position that the mutuality of participant role expectations is indeed an important influence in psychotherapy. He argued that the available evidence indicated that adverse effects follow the disconfirmation of client expectations of the therapist's role. Goldstein also urged that the logical next questions be investigated, such as the endurance of the alleged adverse effects and the malleability of clients' incongruent role expectations.

In summary, the early research in this area constituted an auspicious beginning. It may be due in part to the enthusiastic response elicited by this early work that the hypothesized impact of disconfirmed role expectations has become such a common assumption today. Nevertheless, a review of the efforts since 1962 to validate and extend the early findings indicates that the post-1962 research did not lend itself to the same unequivocal and enthusiastic conclusions. The discussion of this later research is organized around the various types of dependent variables that were used. It summarizes empirical efforts to demonstrate the effects of disconfirmed client role expectations as reflected in client satisfaction, premature termination, outcome, process, and changes in client expectations.

Experimental Research Since 1962

Client Satisfaction

Isard and Sherwood (1964) reported that client satisfaction was not related to the particular interview style of the counselor as long as that style was similar to the client's expectation of how the interview would proceed. Mendelsohn (1964) found that when clients recognized that the therapist was responding differently from their expectations, they rated the interviewer more critically than did clients whose expectations were con-

firmed. Interestingly, some of the clients failed to recognize that the therapist was responding in a style different from the one they expected. Goin, Yamamoto, and Silverman (1965) examined the satisfaction of those clients in their sample who expected advice from their therapists. Of this group, 72% of those who received advice from the therapist reported satisfaction. Only 57% of those who did not receive advice indicated satisfaction with therapy. Severinson (1966) reported that disconfirmed client expectations of the degree of therapist empathy resulted in lesser client satisfaction. The relationship held without regard to the direction of the disconfirmation. Kumler (1969) found a similar relationship between disconfirmed expectations of therapist age and warmth and lesser client satisfaction. It is important to note, however, that the negative effects resulted primarily from cases in which expectations of high therapist warmth were not confirmed. Therapist warmth was positively related to client satisfaction regardless of client expectation.

On the negative side, Cundick (1963) found no significant relationship between the congruence of client–therapist role expectations and satisfaction with psychotherapy as rated by counselor and client. Geller (1965) reported similar findings. Further, Severinson (1966) and Kumler (1969), who found evidence supporting the negative effect on satisfaction of disconfirmed client expectations of empathy and warmth, respectively, both failed to find similar effects of disconfirmed expectations of therapist directiveness. In addition, Klepac (1970), using only the audio portion of the videotapes presented by Kumler, failed to replicate Kumler's finding that disconfirmed expectations of therapist warmth led to lesser client satisfaction. Consistent with Kumler, Klepac also reported no relationship between client satisfaction and disconfirmed expectations for therapist directiveness. Finally, Gladstein's (1969) research suggested that client expectations for the therapist's role and the psychotherapy process were multidimensional. His data indicated that disconfirmation of role expectations resulted in diminished client satisfaction only when none of the several dimensions of expectation was confirmed. Thus, he argued that any one given expectation is not itself an important factor in psychotherapy.

Premature Termination

Other researchers have used premature termination of psychotherapy as the dependent variable. Overall and Aronson (1963) compared client expectations with drop-out rate from therapy. A *dropout* was defined as a client who failed to return for the second interview. The authors found that clients with especially high expectations of an actively supportive, directive therapist dropped out of therapy at a significantly higher rate. They attributed this result to the fact that their therapists tended to see their own role as less directive and less actively supportive. It is unfortunate that the measure of client expectation in this oft cited study was designed without safeguards against a "yes" set by the respondent. The questionnaire was administered orally by a staff social worker, thus making the danger of yes sets especially strong. This threat to internal validity can be used to formulate an alternative explanation for the authors' observation that, in general, clients tended to expect all things from the therapist. On the other hand, enough variance among clients did exist to establish that nonreturners held expectations that were more directive and, therefore, more discrepant from the prevailing therapist expectations at that clinic than were the expectations of those who did return for their second session.

In a similar vein, Borghi (1968) found himself unable to differentiate terminators from continuers in therapy on measures of ego strength, anxiety, and dependency. Subsequent detailed interviews with these clients strongly suggested that early termination of treatment occurred most frequently when these clients held role expectations that were incongruent with those held by the therapists. One of the most commonly held incongruent expectations was that the therapist would give a great deal of advice. Sandler

(1975) reported that the expectations of therapist behavior held by early terminators were more discrepant from their therapists' own role expectations than were the expectations of those who continued in therapy. Premature terminators tended to expect more advice from the therapist and expected the therapist to be more responsible for leading the interview interaction.

Opposing these positive findings, Goin et al. (1965) found no difference in length of stay between those clients whose expectation of advice was met and those whose expectation was not met. Fiester (1974) reported no relationship between disconfirmed client expectations for therapist role and dropping out of psychotherapy. Similarly, Vail (1974) found no relationship between the extent of the discrepancy between clients' and therapists' role expectations of the first interview and continuation in treatment. Finally, Horenstein (1974) measured the extent to which clients thought their expectations of therapist role were confirmed or disconfirmed in the actual interviews. He found no significant relationship between failure to confirm expectations and dropping out of therapy. There was the suggestion of a nonlinear trend in the data, with disconfirmation contributing more to negative effects than confirmation to positive effects.

Psychotherapy Outcome

Several studies have used psychotherapy outcome as the dependent variable. H. Goldstein (1965) reported that describing the potential therapist to the client in terms congruent with the client's expectations of therapist role was quite instrumental in establishing a placebo effect. The effect was even stronger than that produced by positive prognostic expectancy. Dougherty (1973) matched one group of clients with therapists based on a number of variables including their respective orientations to therapy. He found that these "optimally matched" clients had more success in therapy, as rated by their therapists, than did the poorly matched and nonmatched subjects. Unfortunately, it is not easy to interpret these results, since the therapists' perceptions of success may merely reflect their greater comfort with clients who share their expectations of how psychotherapy will proceed. Gulas (1974) also found that clients whose initial role expectations were highly congruent with those held by their therapists demonstrated greater improvement in psychotherapy than those with less congruent expectations.

On the negative side, Volsky, Magoon, Norman, and Hoyt (1965) found no evidence in their data to support the position that clients' expectations about their role, the therapist's role, or other aspects of the therapy process have an important bearing on psychotherapy outcome. In addition, Horenstein (1974) found no relationship between the disconfirmation of client expectations of therapist role and therapy outcome. As with the premature termination data, he did uncover a nonsignificant parabolic trend, such that disconfirmation of expectation contributed more to unsuccessful therapy outcome than confirmation of expectation contributed to successful outcome.

Psychotherapy Process

Research investigating the effects of disconfirmed expectations on the therapy process has produced results that are equally ambiguous. Clemes and D'Andrea (1965) derived their hypothesis from a consideration of an earlier study (Pope & Siegman, 1962) in which the less structured, low-directive therapist style produced greater anxiety in the clients than did the high-directive style. Clemes and D'Andrea postulated that it was not the low-directive therapist style per se that aroused anxiety during the interviews. Rather, they argued, it was the interaction of that style with the clients' expectations for a high-directive style that was the causal agent. To test their hypothesis, Clemes and D'Andrea measured each client's expectations for high–low therapist directiveness and then manipulated the actual interviewer style to which each client was exposed. They found that, independent of the particular interviewer style, clients whose ex-

pectations were not confirmed reported greater anxiety than their counterparts whose expectations were realized. Unfortunately, this self-rating of anxiety was not confirmed by a complementary card-sort measure of anxiety, leaving the results unclear.

Pope, Siegman, Blass, and Cheek (1972) also found effects of disconfirmed expectations on interview process. These writers experimentally induced in one of their two groups the expectation that the second of two interviews would comprise a test interpretation by the therapist. The control group was correctly informed that in the second session the therapist would ask questions soliciting personal information. This sort of manipulation, of course, is very strong. It is akin to a therapist deliberately misrepresenting his or her therapy style to a new client. It did effect differences in interview speech behavior between the two groups. The experimental group was less orally productive and displayed more avoidance in their speech patterns. Pope et al. interpreted these results as support for the hypothesis that the experimental group would exhibit more strain in the oral interaction of the interview. Ziemelis (1974) also included an expectation manipulation as part of his study of the process of the psychotherapy interview. He reported some effects of the expectation manipulation: In general, a client's expectation that he or she would not get the type of therapist desired resulted in negative effects, as reflected in some of the pencil-and-paper measures of the process. This effect was not reflected in the ratings of the actual depth of interaction.

On the negative side, Warren (1973) reported that disconfirmed role expectations did not result in any lower rating of relationship quality or therapeutic conditions as measured by the Barrett-Lennard Relationship Inventory. In addition, Klepac and Page (1974) challenged the generalizability of the results reported by Pope et al. (1972). The former writers summarized certain relevant aspects of an earlier study in which Klepac (1970) studied the effects of not confirming client expectations for the therapist's level of directiveness. In that investigation, each person was interviewed by a therapist whose level of directiveness was manipulated to be high or low, thus confirming or disconfirming the person's expectation. The interview itself was conducted via closed-circuit television. To maintain the picture on the monitor, subjects were required to press a switch at a given rate. Thus, two measures of process were available. First, the amount of switch pressing was recorded and was understood as an index of the reinforcement value of the interviewer. Second, oral productivity was recorded in a manner equivalent to that used by Pope et al. Klepac's results indicated that persons whose expectations for a low-directive therapist were disconfirmed exhibited significantly more switch pressing than did persons in any other groups. Thus, not only were there no negative effects of the failure to confirm client expectations but there was a distinctly positive effect of the failure to confirm expectations for a low-directive therapist. Finally, Horenstein (1974) also included a measure of interview process: resistance to psychotherapy. As with the drop-out and outcome variables, he found no relationship between confirmation of expectation and degree of client resistance exhibited, but he did note the parabolic trend that was also reported for the premature termination and therapy outcome variables.

Change in Expectation

Another type of dependent variable used in role expectation research has been the extent and direction of change in the expectation itself as a result of its confirmation or disconfirmation. If role expectations are indeed an important influence in psychotherapy, it is hypothesized that they should be rather strongly held. Kumler (1969) reported that client expectations were relatively stable in the face of disconfirmation. However, clients did tend to change their expectations in the direction of very warm therapists, regardless of original expectation. Sandler (1975) noted that there was a reduction of dissimilarity in role expectation as therapy continued but that clients who eventually dropped out of therapy exhibited

a much lesser tendency to change expectations in the direction of actual therapist behavior.

On the other hand, Cundick (1963) found that client and therapist role expectations became significantly more congruent as counseling progressed. Klepac (1970) was unable to replicate Kumler's results using only the audio portion of Kumler's videotape. Klepac reported that expectations changed in the direction of the therapist's actual style. He replicated his own findings in a second study (Klepac, 1970) in which he reported that client expectations regarding therapist directiveness were quite fluid, again conforming to the assigned therapist's style. Finally, Gulas (1974) also found that congruence of client–therapist role expectations increased as therapy progressed.

Shaping Role Expectation

One other approach has been tried in the attempt to demonstrate the hypothesized relationship between disconfirmed role expectations and negative consequences. The logic underlying this research might be reconstructed as follows: If it is important that client role expectations be confirmed, then clients whose role expectations have been systematically shaped toward congruence with actual therapist style should exhibit more positive consequences. Albronda, Dean, and Starkweather (1964) instructed psychiatric social workers to use preliminary client contacts to help clients form accurate expectations regarding the nature of psychotherapy and the roles of the participants. The authors reported that clients who received this treatment dropped out of therapy at a lesser rate and exhibited improved outcome. Hoehn-Saric et al. (1964) reported that role induction interviews significantly improved client outcome in psychotherapy. The role induction interview consisted of a brief session structured to develop accurate expectations of the psychotherapy process. These results were replicated in a later study by Schonfield, Stone, Hoehn-Saric, Imber, and Pande (1969).

Related manipulations have also proven successful. Mosby (1972) used a procedure in which he informed therapists of existing differences between their own role expectations and the role expectations held by their clients. The therapists were told to try to modify their clients' expectations early in therapy to conform more closely with their own. Clients in the experimental condition changed their expectations of the therapist's role more quickly and dropped out of therapy at a lesser rate than did the controls. Heilbrun (1972) attempted to modify client expectations by using a brief pretherapy session in which clients were instructed that counselors of various styles could be equally effective and that the client should try to adapt his or her expectations to the counselor's style to obtain maximum benefit. Heilbrun found that clients characterized by low readiness for therapy demonstrated a lower drop-out rate when they received the pretherapy briefing than when they did not. However, high-readiness clients were not differentially affected by the briefing process.

Contrary to these uniformly positive results, Venema (1972) was unable to demonstrate the relationship between expectations shaped before therapy and positive therapeutic consequences. An important difference between Venema's study and the studies that reported positive findings is that Venema did not use live interaction in the change-of-expectation process. Instead he used a videotaped manipulation of inappropriate role expectations. In using a procedural check, Venema insured that the group exposed to the videotape entered therapy with significantly more accurate expectations of the therapist's role than did the control group. The experimental group also experienced fewer role expectation disconfirmations during the initial interview. Nevertheless, Venema found no relationship between role expectation disconfirmation and attrition. It may be that it was not the modification of expectations at all that effected the positive outcomes in the previous studies. Rather, it may be that the extra personal attention afforded the clients who received the role induction interviews—the one most important condition that discriminates between the

earlier studies and Venema's—accounted for their more positive results. This alternative interpretation is supported by Fernbach's (1975) failure to effect more positive results with clients who received a written clarification of therapist and client roles. In addition, Orenstein (1974) found no statistically significant effects of a role preparation tape and an attraction induction message on subjects' perceptions of the relationship and of the psychotherapist. It might be fruitful to compare in a single study the separate and interactive effects of extra attention and extra information on client expectation and on psychotherapeutic process and outcome.

In summary, a comprehensive review of the available literature suggests considerable ambiguity regarding the validity of the hypothesis that disconfirmed role expectations result in negative consequences. Research since 1962 has not clearly supported the enthusiastic inferences drawn from the early speculative and experimental literature, nor does the hypothesis warrant the assumption of *proven* that it has so often enjoyed in the current literature. The "box score" may be summarized as follows. Single studies are counted more than once if they used two or more distinct classes of dependent variables or two or more clearly distinct independent variables.

Among studies that used satisfaction as the dependent variable, five (45%) supported the hypothesized relationship and six (55%) did not support it. Among studies that used premature termination from psychotherapy as the dependent variable, three (43%) supported the hypothesized relationship and four (57%) did not. Three studies (60%) that used outcome as the dependent variable supported the hypothesized relationship and two (40%) did not. Studies of the effect of disconfirmed expectations on psychotherapy process were evenly split. Three supported the hypothesized relationship and three did not. Among studies that measured the tenacity with which disconfirmed expectations were held, two (33%) found expectations to be held quite strongly. On the other hand, four studies (67%) found expectations to be rather fluid, changing in the direction of the assigned therapist. The former results support the hypothesized relationship; the latter do not. Finally, five studies (62%) found that pretherapy modification of expectations resulted in more positive consequences; three others (38%) found no effects of pretherapy induction efforts.

The empirical foundation for the hypothesized negative effects of disconfirmed role expectations is imbued with a great deal of ambiguity. In only one type of empirical investigation of this problem—the role induction strategy—is there any hint of a predominance of studies in favor of the expectation hypothesis. Even in this research, the interpretations are confounded by the failure of most of the positive studies to separate the effects of extra personal attention from those that resulted from changes in role expectations. Overall, 21 studies (49%) supported the hypothesized relationship; 22 studies (51%) did not. It would be difficult to find a more even split in a research area. The ambiguity is highlighted even more when one remembers the reluctance of so many editors and reviewers to judge favorably those manuscripts that report nonsignificant results and the consequent reluctance of many researchers to submit such results. Clearly, the conclusion that it is important to meet client role expectations in psychotherapy does not deserve the sustained support that it has received. It must be removed from its status as *demonstrated* in the common wisdom of psychology.

Assessment of the Literature

This section considers the contribution of several factors to the state of ambiguity regnant in the role expectation literature. Implicit in the discussion of these factors are implications for improvements in future research design to clarify the importance of client role expectations in psychotherapy.

Klepac and Page (1974) addressed themselves to the problem of contradictory findings between their report and that of Pope et al. (1972). They argued that one factor that contributes to ambiguity in research on role expectations is the tendency of researchers in the area to study imprecisely defined

or globally assessed expectations. Venzor, Gillis, and Beal (1976) found a clear lack of correspondence between subjects' responses to an adjective checklist of preferred therapist characteristics and a quasi-behavioral index of preferred therapist style. They suggested that future studies use measures that are as close as possible to a behavioral representation of the therapist's behavior. Only a few studies have assessed role expectations with video, audio, or written examples of the behavior in question. Further, some investigators have used instruments for the assessment of expectation that use open-ended response formats, which permit considerable subjective bias in scoring. In one case, the items were objectively written, but the questions were asked orally of the patient by an intake worker. Not surprisingly, these investigators reported that the patients expected almost everything from the therapist; that is, they answered *yes* to almost every question that began "Do you expect . . . ?" Methodologically, one must suspect that a positive response bias may have resulted. In other situations, authors have used relatively brief instruments to assess expectations on a number of dimensions. Since only a few items could be applied to each dimension, the question of reliability must be raised. Finally, some researchers have used instruments of adequate characteristics but have elected to measure expectancy for therapist characteristics that are by definition rather broad in nature, for example, therapist personality.[1]

It is this multifaceted tendency toward imprecision that Klepac and Page (1974) held to be a major cause of the inconsistency of experimental findings in the research on client role expectations. They proposed that future research attempt the twofold task of (a) establishing which are the most relevant basic dimensions along which client role expectations may be described and (b) determining the nature of the effects, if any, of failing to meet client expectations along these various dimensions. This paradigm has a real potential for building a more solid empirical basis on which to make judgments regarding the actual relevance of client role expectations to psychotherapy.

A second problem area in the research on role expectations has been the ambiguous definition of the term *expectation*. In the precise sense in which it was originally used by Kelly (1955) in his theory of personality and psychotherapy, and by Apfelbaum (1958) in the classic study of role expectations, expectation was clearly defined as the *anticipation* of some event. There was the implication that the anticipation is held with some degree of certainty. Klepac (1970) and Pope et al. (1972) are examples of later researchers who have been careful to define expectation for their subjects as this same sense of anticipation. However, most researchers have not been so careful in drawing the distinction, for their subjects or for themselves, between this definition of expectation and the alternative, competing connotation that can be attached to the same word. In this alternative connotation, the term expectation can carry the implication that the expector is due, has the right to, or demands that which is expected. This meaning of the term suggests that the person to whom the expectation is extended has some obligation to meet that expectation. To say to another person "I expect you to be forceful with your boss!" clearly may be understood to imply more than mere anticipation. In truth, such a statement uses *expect* to communicate a desire that such behavior will be forthcoming. More precisely, it is a *preference* that some event should occur.

It seems straightforward to state that these two concepts—expectation as anticipation and expectation as preference—are sufficiently different aspects of human cognition to warrant distinct treatment. Nevertheless, most researchers in the area of role expectations have not dealt specifically with this issue (cf. Rosen, 1967). The majority of authors have simply neglected the problem,

[1] The authors are grateful to an anonymous reviewer who noted that expectations of the therapist's personal qualities and expectations of the therapist's behavior may be differentially determined and may respond differently to disconfirmation. This conceptual distinction has not often been made and may warrant closer consideration in future studies.

apparently leaving subjects to interpret expectation as they wished. The interpretation of their results is not clear, therefore, and the variance among studies is not surprising. Some other authors, while indicating that they were studying expectations, precisely defined their independent variable in terms of client preference for some therapist behavior. For example, one investigator's instruction to describe one's ideal therapist is clearly not likely to have elicited client expectations (anticipations) of therapist behavior. Thus, the imprecision with which expectation has been defined is a second factor in the inconsistency of the research results. The series of studies in this area may have intermittently manipulated one of two distinctly different independent variables. The definitional problem is compounded by the fact that few researchers have stayed with this topic for more than one study. This fluctuation has added considerably to the already large variance in what precisely is being assessed as the independent variable.

The third factor discussed in the present article cannot be so definitively stated. It is presented for consideration by researchers in the area. It is speculated that expectation and preference may not only be related, but may be related hierarchically, such that studies of client expectation cannot be unequivocally understood without the simultaneous investigation of preference.

As early as 1955, Shaw noted that client role expectations might be disconfirmed in a desired direction as well as in an undesired direction. Helson (1959, 1964) developed a sophisticated statement of a similar position. He developed this statement as an alternative to the earlier hypothesis regarding the effects of disconfirmed expectation that was presented by McClelland, Atkinson, Clark, and Lowell (1953) in their elaboration of the achievement motive. Helson's hypothesis regarding the effects of disconfirmed expectation was a bipolar one. He postulated that when the expectation under investigation is one that embodies an affective component, then the affective and motivational consequences associated with the disconfirmation of the expectation arise as a function of the *direction,* as well as the intensity, of the discrepancy.

Helson did not question the adequacy of McClelland et al.'s model for explaining and predicting reactions to the disconfirmation of expectations that are primarily sensory in nature. However, he hypothesized that expectations with an affective or aesthetic component are markedly different in nature from expectations of a purely sensory event. He reasoned that the former types of expectations could be ranked along a continuum ranging across levels of desirability, from most preferred to least preferred, passing through a zone of indifference. Therefore, Helson deduced that given some particular affectively toned expectation, a discrepancy from that expectation in one direction along the desirability continuum would elicit reactions very different in quality from those elicited by a discrepancy in the opposite direction. If the event that actually occurs is more desirable than the expected event, the result will be positive affect and approach motivation. Conversely, if the actual event is less desirable than the expected event, the result will be negative affect and avoidance motivation. Further, Helson also predicted that as the actual event extends farther from the expected event along the desirability continuum, it will elicit increasingly strong positive or negative reactions. In summary, it is clear that Helson considered preference to be a more basic variable than expectation, undergirding it so to speak, and one that must necessarily be known if one is to predict the nature of a given person's response to disconfirmation of an expectation.

This bipolar position stands in contrast to the unidimensional theory of reactions to disconfirmed expectations developed by McClelland et al. (1953). The unidimensional position implicitly or explicitly underlies most investigators' hypotheses that disconfirmation of client role expectations in psychotherapy will negatively influence process or outcome. The unidimensional position holds that the resultant affective and motivational states arise solely as a function of the extent of the discrepancy between the actual event and the expected event. A slight dis-

crepancy in either direction will elicit reactions slightly more positive than will no discrepancy. Beyond that point, the larger the discrepancy, the more negative the reaction.

If the unidimensional theory of expectation disconfirmation is adequate, preference should be considered a noninteracting variable with expectation. On the other hand, if the bipolar theory provides a more adequate explanation for affectively and aesthetically toned expectations, then preference must be considered to be a variable inextricably intertwined with expectation and integral to an understanding of reaction to disconfirmation of expectation. The bipolar theory predicts that the preference modifies the quality as well as the quantity of the reaction to disconfirmed expectation.

Block (1964) undertook an empirical comparison of the two theories of reaction to disconfirmed expectation as they apply to client role expectation in psychotherapy. Experienced raters were trained to note all client remarks that expressed a felt discrepancy between the actual therapy experience and the prior expectation. These remarks were categorized in terms of the size and the direction (more preferred – less preferred) of the discrepancy. Manifest client affect and inferred secondary motivational states were also recorded. Approach motives in psychotherapy were operationalized to include such behaviors as taking initiative in talking, presenting new associations and dreams, making productive use of silence, and exhibiting actions and movements that reflect deeper involvement in the psychotherapy. Avoidance motives were operationalized to include such behaviors as evasions, forgetting dreams, angry silences, coming late, threatening termination, and withdrawing to shallower involvement in psychotherapy. The results clearly supported the bipolar position's predictions. Affective responses and inferred secondary motives varied with the direction of the discrepancy. There was no effect of the size of the discrepancy alone. Thus, the question of whether the client's expectations were confirmed or disconfirmed is too simplistic when asked alone. One must also ask whether the person wanted or did not want what he or she expected.

Although Block's study was reported some time ago, and recently was cited in some detail by Meltzoff and Kornreich (1970), researchers in the area of role expectations in psychotherapy have seemingly not grappled with its implications. It appears at least to us that role expectation research has continued to be based on the unidimensional position, although without such specific conceptual elaboration of the basic assumptions on which the hypotheses are based. We hypothesize that inattention to the bipolar position and the resulting failure to account for the preference variable are two of the factors that have led to the currently highly ambiguous state of the research on disconfirmation of expectation in psychotherapy.

Summary

Early research findings lent support to previous speculation regarding the importance of meeting client role expectations in psychotherapy. The beginnings of this research effort were auspicious. Reviewers of the expectation research offered strong statements regarding the importance of role expectation in psychotherapy. Such statements were met with enthusiastic acceptance, in part because they appeared to herald an alliance of social psychology and psychotherapy research. The belief that negative effects result from the failure to confirm client role expectations was quickly and easily accepted into the common wisdom of psychology. This belief is typically considered to be demonstrated by the research and is usually cited as such without encountering serious challenge. A review of the available empirical literature since 1962 revealed that the validity of the disconfirmed expectations – negative effects hypothesis has not been established with certainty. The research was almost evenly divided in terms of support for and lack of support for the hypothesis.

The ambiguous state of the research was discussed in terms of problems in the designs and conceptualizations that have been used. Namely, it was noted that (a) the opera-

tionalization of the independent variable has often not been adequately precise or reliable, (b) the definition of expectation has usually not been clearly specified for the reader or subject, and (c) the theoretical position that has implicitly undergirded almost all the research on role expectation in psychotherapy may not be appropriate for the kinds of affectively toned expectations that are involved in psychotherapy. In conclusion, it would be most appropriate to approach the subject of role expectations more cautiously for the present. Continued research will be necessary to evaluate the true impact of disconfirmed expectations in psychotherapy. This research should incorporate procedures to counter the types of problems that have limited past work. In the interim, it may be unwarranted to continue to publish the many limited reports that describe the various role expectations of different client groups. It may not be important, in a functional sense, what their expectations are. Furthermore, theses based on the so-called established relationship of disconfirmed expectations and negative effects in psychotherapy should be reexamined in light of the fact that their relationship is not as clearly understood as has been supposed.

References

Albronda, H., Dean, R., & Starkweather, J. Social class and psychotherapy. *Archives of General Psychiatry*, 1964, *10*, 276–283.

Apfelbaum, B. *Dimensions of transference in psychotherapy*. Berkeley: University of California Press, 1958.

Baekeland, F., & Lundwall, L. Dropping out of treatment: A critical review. *Psychological Bulletin*, 1975, *82*, 738–783.

Baldwin, B. Self-disclosure and expectations for psychotherapy in repressors and sensitizers. *Journal of Counseling Psychology*, 1974, *21*, 455–456.

Begley, C., & Lieberman, L. Patient expectations of therapists' techniques. *Journal of Clinical Psychology*, 1970, *26*, 112–116.

Bent, R., Putnam, D., Kiesler, D., & Nowicki, S., Jr. Expectancies and characteristics of outpatient clients applying for services at a community mental health center facility. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 280.

Biddle, B. An application of social expectation theory to the initial interview (Doctoral dissertation, University of Michigan, 1957). *Dissertation Abstracts*, 1958, *19*, 186. (University Microfilms No. 58-1377)

Block, W. A preliminary study of achievement motive theory as a basis of patient expectations in psychotherapy. *Journal of Clinical Psychology*, 1964, *20*, 268–271.

Borghi, J. Premature termination of psychotherapy and patient-therapist expectations. *American Journal of Psychotherapy*, 1968, *22*, 460–473.

Caine, T., Wijesinghe, B., & Wood, R. Personality and psychiatric treatment expectancies. *British Journal of Psychiatry*, 1973, *122*, 87–88.

Chance, E. Mutual expectations of patients and therapists in individual treatment. *Human Relations*, 1957, *10*, 167–178.

Chance, E. *Families in treatment*. New York: Basic Books, 1959.

Clemes, S., & D'Andrea, V. Patients' anxiety as a function of expectation and degree of initial interview ambiguity. *Journal of Consulting Psychology*, 1965, *29*, 397–404.

Cundick, B. The relation of student and counselor expectations to rated counseling satisfaction (Doctoral dissertation, Ohio State University, 1962). *Dissertation Abstracts*, 1963, *23*, 2983–2984. (University Microfilms No. 63-0044)

Danskin, D. Roles played by counselors in their interviews. *Journal of Counseling Psychology*, 1955, *2*, 22–27.

Dougherty, F., III. Patient-therapist matching: An empirical approach toward the improvement of psychotherapy outcome (Doctoral dissertation, Vanderbilt University, 1972). *Dissertation Abstracts International*, 1973, *33*, 6074B. (University Microfilms No. 73-14505)

Fernbach, R. Preparation of clients for individual psychotherapy using a written document to orient expectations and indicate appropriate behaviors (Doctoral dissertation, Ohio University, 1974). *Dissertation Abstracts International*, 1975, *35*, 6092B–6093B. (University Microfilms No. 75-11963)

Fiester, A. Pre-therapy expectations, perception of the initial interview and early psychotherapy termination: A multivariate study (Doctoral dissertation, Miami University, 1974). *Dissertation Abstracts International*, 1974, *35*, 1907B. (University Microfilms No. 74-21729)

Frank, J., Gliedman, L., Imber, S., Nash, E., & Stone, A. Why patients leave psychotherapy. *Archives of Neurological Psychiatry*, 1957, *77*, 283–299.

Garfield, S., & Wolpin, M. Expectations regarding psychotherapy. *Journal of Nervous and Mental Disease*, 1963, *137*, 353–362.

Geller, M. Client expectations, counselor role perception, and outcome of counseling (Doctoral dissertation, University of California, Berkeley, 1965). *Dissertation Abstracts*, 1965, *26*, 4073. (University Microfilms No. 65-13489)

Gladstein, G. Client expectations, counseling experience, and satisfaction. *Journal of Counseling Psychology*, 1969, *16*, 476–481.

Goin, M., Yamamoto, J., & Silverman, J. Therapy congruent with class-linked expectations. *Archives of General Psychiatry,* 1965, *13,* 133–137.

Goldstein, A. Participant expectancies in psychotherapy. *Psychiatry,* 1962, *25,* 72–79. (a)

Goldstein, A. *Therapist and patient expectancies in psychotherapy.* New York: Macmillan, 1962. (b)

Goldstein, H. Placebo psychotherapy and change in anxiety, mood, and adjustment (Doctoral dissertation, University of Florida, 1965). *Dissertation Abstracts,* 1965, *26,* 1775. (University Microfilms No. 65-9602)

Gulas, I. Client-therapist congruence in prognostic and role expectations as related to client's improvement in short-term psychotherapy (Doctoral dissertation, Ohio University, 1974). *Dissertation Abstracts International,* 1974, *35,* 2430B. (University Microfilms No. 74-23852)

Hankoff, L., Englehardt, D., & Freeman, N. Placebo in schizophrenic outpatients. *Archives of General Psychiatry,* 1960, *2,* 33–42.

Heilbrun A. Effects of briefing upon client satisfaction with the initial counseling contact. *Journal of Consulting and Clinical Psychology,* 1972, *38,* 50–56.

Heine, R., & Trosman, H. Initial expectations of the doctor-patient interaction as a factor in continuance in psychotherapy. *Psychiatry,* 1960, *23,* 275–278.

Heitler, J. Preparatory techniques in initiating expressive psychotherapy with lower-class, unsophisticated patients. *Psychological Bulletin,* 1976, *83,* 339–352.

Helson, H. Adaption-level theory. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 1). New York: McGraw-Hill, 1959.

Helson, H. *Adaption-level theory.* New York: Harper & Row, 1964.

Hoehn-Saric, R., et al. Systematic preparation of patients for psychotherapy: I. Effects on therapy behavior and outcome. *Journal of Psychiatric Research,* 1964, *2,* 267–281.

Horenstein, D. The effects of confirmation or disconfirmation of client expectations upon subsequent psychotherapy (Doctoral dissertation, University of Kansas, 1973). *Dissertation Abstracts International,* 1974, *34,* 6211B. (University Microfilms No. 74-12575)

Isard, E., & Sherwood, E. Counselor behavior and counselor expectations as related to satisfactions with counseling interview. *Personnel and Guidance Journal,* 1964, *42,* 920–921.

Jacobs, M., Muller, J., Anderson, J., & Skinner, J. Therapeutic expectations, premorbid adjustment, and manifest distress level as predictors of improvement in hospitalized patients. *Journal of Consulting and Clinical Psychology,* 1972, *39,* 455–461.

Kamm, R., & Wrenn, C. Client acceptance of self information in counseling. *Educational and Psychological Measurement,* 1950, *10,* 32–42.

Kelly, G. *The psychology of personal constructs* (Vol. 2). New York: Norton, 1955.

Klepac, R. An experimental analogue of psychotherapy involving "client" behavior as a function of confirmation and disconfirmation of expectations of "therapist" directiveness (Doctoral dissertation, Kent State University, 1969). *Dissertation Abstracts International,* 1970, *30,* 5690B–5691B. (University Microfilms No. 70-11348)

Klepac, R., & Page, H. Discrepant role expectations and interviewee behavior: A reply to Pope, Siegman, Blass, and Cheek. *Journal of Consulting and Clinical Psychology,* 1974, *42,* 139–141.

Kumler, M. Client expectations of therapist role: Relationship to initial commitment in a psychotherapy analogue (Doctoral dissertation, Kent State University, 1968). *Dissertation Abstracts,* 1969, *29,* 4848B–4849B. (University Microfilms No. 69-9561)

Lennard, H., & Bernstein, A. *The anatomy of psychotherapy.* New York: Columbia University Press, 1960.

Lorion, R. Patient and therapist variables in the treatment of low-income patients. *Psychological Bulletin,* 1974, *81,* 344–354. (a)

Lorion, R. Social class, treatment attitudes, and expectations. *Journal of Consulting and Clinical Psychology,* 1974, *42,* 920. (b)

McClelland, D., Atkinson, J., Clark, R., & Lowell, L. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.

McGowan, J. Client anticipations and expectancies as related to initial interview performance and perceptions (Doctoral dissertation, University of Missouri—Columbia, 1954). *Dissertation Abstracts,* 1955, *15,* 228–229. (University Microfilms No. 00-10120)

McNair, D., & Lorr, M. An analysis of professed psychotherapeutic techniques. *Journal of Consulting Psychology,* 1964, *28,* 265–271.

Meltzoff, J., & Kornreich, M. *Research in psychotherapy.* New York: Atherton Press, 1970.

Mendelsohn, R. The effects of cognitive dissonance and interview preference upon counseling-type interviews (Doctoral dissertation, University of Michigan, 1963). *Dissertation Abstracts,* 1964, *24,* 2987–2988. (University Microfilms No. 64-860)

Mosby, R. Alteration of clients' expectations about counseling in the direction of client-counselor mutuality by means of an experimental intervention procedure (Doctoral dissertation, University of Texas at Austin, 1971). *Dissertation Abstracts International,* 1972, *33,* 446B–447B. (University Microfilms No. 72-19635)

Orenstein, L. Pre-therapy role preparation and attraction induction: An experimental analogue (Doctoral dissertation, Kent State University, 1973). *Dissertation Abstracts International,* 1974, *34,* 3505B. (University Microfilms No. 73-32353)

Overall, B., & Aronson, H. Expectations of psychotherapy in patients of lower socio-economic class. *American Journal of Orthopsychiatry,* 1963, *33,* 421–430.

Patterson, C. Client expectations and social conditioning. *Personnel and Guidance Journal,* 1958, *36,* 136–138.

Pope, B., & Siegman, A. The effect of therapist verbal activity level and specificity on patient productivity and speech disturbance in the initial interview. *Journal of Consulting Psychology,* 1962, *26,* 489.

Pope, B., Siegman, A., Blass, T., & Cheek, J. Some effects of discrepant role expectations on interviewee verbal behavior in the initial interview. *Journal of Consulting and Clinical Psychology,* 1972, *39,* 501–507.

Rosen, A. Client preference: An overview of the literature. *Personnel and Guidance Journal,* 1967, *45,* 785–789.

Sandler, W. Patient-therapist dissimilarity of role expectations related to premature termination of psychotherapy with student therapists (Doctoral dissertation, City University of New York, 1975). *Dissertation Abstracts International,* 1975, *35,* 6111B–6112B. (University Microfilms No. 75-12691)

Schonfield, J., Stone, A., Hoehn-Saric, R., Imber, S., & Pande, S. Patient-therapist convergence and measures of improvement in short-term psychotherapy. *Psychotherapy: Theory, Research and Practice,* 1969, *6,* 267–272.

Seeman, J. An investigation of client reactions to vocational counseling. *Journal of Consulting Psychology,* 1949, *13,* 95–104.

Severinson, J. Client expectation and perception of the counselor's role and their relationship to client satisfaction. *Journal of Counseling Psychology,* 1966, *13,* 109–112.

Shaw, F. Mutualities and up-ending expectancies in counseling. *Journal of Counseling Psychology,* 1955, *2,* 241–247.

Skinner, K., & Anderson, G. Personality and attitude characteristics associated with therapy readiness. *American Psychologist,* 1959, *14,* 367–377.

Thomas, E., Polansky, N., & Kounin, J. The expected behavior of a potentially helpful person. *Human Relations,* 1955, *8,* 165–174.

Tinsley, H., & Harris, D. Client expectations for counseling. *Journal of Counseling Psychology,* 1976, *23,* 173–177.

Vail, A. Dropout from psychotherapy as related to patient-therapist discrepancies, therapist characteristics, and interaction in race and sex (Doctoral dissertation, Fordham University, 1974). *Dissertation Abstracts International,* 1974, *35,* 2452B. (University Microfilms No. 74-25087)

Venema, J. The effects of expectancy training, commitment, and therapeutic conditions upon attrition from outpatient psychotherapy (Doctoral dissertation, Fuller Theological Seminary, 1970). *Dissertation Abstracts International,* 1972, *32,* 6664B–6665B. (University Microfilms No. 72-15871)

Venzor, E., Gillis, J., & Beal, D. Preference for counselor response styles. *Journal of Counseling Psychology,* 1976, *23,* 538–542.

Volsky, T., Jr., Magoon, T., Norman, W., & Hoyt, D. *The outcomes of counseling and psychotherapy: Theory and research.* Minneapolis: University of Minnesota Press, 1965.

Wallach, M., & Strupp, H. Psychotherapists' clinical judgments and attitudes toward patients. *Journal of Consulting Psychology,* 1960, *24,* 316–323.

Warren, B. Client expectations and the client-counselor relationship in a counseling analogue. JSAS *Catalog of Selected Documents in Psychology,* 1973, *3,* 131. (Ms. No. 492)

Ziemelis, A. Effects of client preference and expectancy upon the initial interview. *Journal of Counseling Psychology,* 1974, *21,* 23–30.

Received October 28, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 276–296

Taste Aversion and the Generality of the Laws of Learning

A. W. Logue
Harvard University

Results from research on aversions acquired through the pairing of ingested substances with illness have recently been used to challenge the assumption that there are laws of learning that hold across different species and tasks. The taste aversion literature is selectively reviewed and compared with data from traditional experiments in order to evaluate this challenge. Areas sufficiently documented or controversial to warrant inclusion are associative fluidity, conditioned stimulus and unconditioned stimulus characteristics, temporal relationships, obtaining and processing information, and age differences. The conclusion is that in no instance are different principles required to describe taste aversion and traditional learning. In some cases large parametric differences between the two research areas are apparent. It is suggested that at the present time it is not necessary to dispense with the notion of general laws of learning.

An increasingly prevalent position in psychology is that more attention must be paid to previously ignored species-specific and task-specific differences in the general laws of learning and that a reassessment of the field is necessary (see, e.g., Bolles, 1973; Breland & Breland, 1961; Hinde, 1973; Lockard, 1971; Schwartz, 1974; Shettleworth, 1972a; Staddon & Simmelhag, 1971; Weisman, 1977). Two widely cited articles in *Psychological Review* (Rozin & Kalat, 1971; Seligman, 1970) dealt extensively with this subject.

Seligman focused on the number of trials or amount of information needed for learning to occur, a dimension he called preparedness (for similar concepts see Capretta's, 1961, stimulus relevance and Thorndike's, 1932, belongingness). According to Seligman, previous psychological research has assumed that all associations are equally prepared; in fact, the number of trials required for learning to take place can vary a great deal. The degree and direction of variation are a function of the organism's inherited associative apparatus, acquired throughout a long evolutionary history. Seligman believes that contraprepared associations, learning that takes a great many trials to occur, may follow different laws than prepared associations, which take very few trials to occur. The characteristics of learning described by these laws might include "acquisition . . . resistance to extinction; maximum delay of reinforcement, flatness of generalization gradient" (Seligman & Hager, 1972, p. 5).

Rozin and Kalat alternatively maintained that there is no a priori reason to expect degree of preparedness to be predictive of other laws of learning. Their position was that laws of learning for each species and each learning situation evolve in such a way as to maximize survival; two learning tasks equivalent in preparedness may or may not be equivalent in other aspects of learning, such as maximum delay of reinforcement.

Supported by Grant MH-15494 from the National Institute of Mental Health to Harvard University, this article was written while the author was a National Science Foundation Graduate Fellow. It is part of a dissertation submitted to Harvard University in partial fulfillment of the requirements for the PhD degree. Special thanks are due to I. Shrank for his help in all stages of the article's preparation. The continued technical assistance of V. Koster, and comments on previous drafts by P. de Villiers, G. Heyman, R. Mansfield, J. Mazur, P. Rozin, and the editorial staff of *Psychological Bulletin*, are also greatly appreciated.

Requests for reprints should be sent to A. W. Logue, who is now at the Department of Psychology, State University of New York, Stony Brook, New York 11794.

Although Seligman and Rozin and Kalat differ in their beliefs about the manner in which evolution influences the laws of learning, both their views emphasize the importance of this influence. Both stress the necessity of modifying or removing existing general laws of learning to accommodate adaptive specializations in learning.

The area of psychology that has received the most attention, attack, and theoretical speculation with respect to these issues is feeding behavior, specifically, the learning of taste aversions. For many years taste aversion acquisition was known as the bait-shy phenomenon. Barnett (1975) has carefully described how rats, who consume their food in discrete meals, will briefly sample a new food (the bait), wait, and if they get ill, subsequently avoid that food.

Bait shyness was introduced into the laboratory by Garcia and his colleagues (Garcia, Kimeldorf, & Koelling, 1955). They discovered that when a gastrointestinal-like illness, produced in rats by irradiation, was paired with food, the food became aversive (Garcia, Kimeldorf, & Hunt, 1961). Eventually these data came to the attention of the psychological community (Seligman & Hager, 1972), and an onslaught of taste aversion experiments began (Riley & Clarke, 1977). Several procedural refinements were made, including the substitution in many studies of injections of the poison lithium chloride (LiCl) for irradiation, since LiCl is easier to administer (Nachman & Ashe, 1973). The convention in the taste aversion literature has been to use classical conditioning terminology, with the taste as the conditioned stimulus (CS) and the illness as the unconditioned stimulus (US) (Garcia, McGowan, & Green, 1972). This convention is followed here.

The taste aversion experiments appeared to demonstrate that there were many differences between taste aversion and traditional laws of learning. In particular, acquisition of taste aversions despite extremely long delays between the CS and the US (Garcia, Ervin, & Koelling, 1966) and a proclivity for animals to associate tastes and not other stimuli with illness (Garcia & Koelling, 1966) were noted. The implications of these findings were widely discussed, with some authors taking the view that the data could be incorporated into existing, or somewhat modified, general laws of learning (Krane & Wagner, 1975; Mackintosh, 1974; Revusky, 1977d; Revusky & Garcia, 1970; Testa & Ternes, 1977) and others feeling that more extensive revision of the laws was necessary (Kalat, 1977; Shettleworth, 1972a; Zahorik & Houpt, 1977; see Rozin, 1977, for a more detailed analysis of the reactions to the anomalous taste aversion findings).

Two problems were inherent in these discussions. First, the traditional laws of learning used for comparison were not clearly specified. Seligman and Garcia (Garcia, Hankins, & Rusiniak, 1974; Seligman, 1970; Seligman & Hager, 1972) referred to the equipotentiality and contiguity assumptions of Thorndike's (1911/1965) law of effect as principles that have been considered to hold in most situations and in most higher species. The former assumption is directly opposed to the concept of preparedness; it states that all stimuli and responses are equally associable. The latter declares that learning will only occur if the stimuli and the responses involved are in close temporal contiguity. Nevertheless, exactly what qualifies as equivalence of associability is never stated; close temporal contiguity is equally difficult to define. Seligman hinted at additional laws, those regarding generalization, extinction, partial reinforcement, and so forth, but these are also never defined. This is not a problem unique to taste aversion theorists, but a statement of the general condition of learning theory. No list of equations presently exists that accurately describes all that we know about the learning process.

Even if we had a list of the laws of learning, another problem would remain. What constitutes an exception to a general law of learning? As Kalat (1977) has pointed out, to learn more easily does not necessarily mean to learn differently. Bitterman (1975) and Hull (1945) have made a distinction between qualitative and quantitative differences in learning with respect to laws that vary between species. These categories can also be employed to assist in the study of laws that may vary within a species as a function of the stimuli and responses involved, as may be the

case with prepared versus unprepared associations. The quantitative–qualitative question boils down to "whether the performance of all animals can be deduced from a common set of principles or whether different principles are necessary" (Bitterman, 1975, p. 707). Thus, if one were attempting to identify merely quantitative differences between taste aversion learning and the general laws of learning, one would look for cases in which the same equations could be used, only with different parameter values. What seems to be an intuitively more fundamental, qualitative difference would necessitate the use of different equations to describe the acquisition of taste aversions and more traditional tasks (Herrnstein, 1977; see Hull, 1945, for a detailed, though hypothetical, example of this method of analysis).

A moment's consideration of the qualitative–quantitative distinction reveals it to be somewhat arbitrary when carried to extremes. A large change in the value of a parameter is dubbed a quantitative change, while the removal of a term whose value is already near zero constitutes a qualitative change. In other words, the continuum of small to large differences in laws of learning does not easily split into two pieces, quantitative and qualitative. However, this dichotomy is useful for investigating and categorizing differences between taste aversion and traditional learning.

This article reviews and compares results from taste aversion and more traditional learning experiments in an attempt to assess the generality of accepted laws of learning. Clearly this presents some difficulties. At times it is possible to determine for a given situation the consensus with respect to how learning should take place (a general law of learning). In these cases the qualitative–quantitative distinction may prove helpful. At other times, the effort to differentiate between principles that describe taste aversion and traditional learning seems no more than shadowboxing because neither side is able to clearly define its stance. In any case, as a result of the multitude of claims and counterclaims that have been made about taste aversion learning and consequently about the generality of learning principles, a close examination of some of the data should be instructive.

Because of the extensiveness of the taste aversion literature, only well-documented areas are discussed and representative studies cited. Emphasis is placed on controversial and unusual findings. Readers interested in obtaining additional information from the taste aversion literature should see the almost 700 references compiled by Riley and his colleagues (Riley & Baril, 1976; Riley & Clarke, 1977). The areas discussed here are associative fluidity, CS and US characteristics, temporal relationships, obtaining and processing information, and age differences.

A comparison of the physiological bases of taste aversion and traditional tasks has not been included for three reasons. First, there has been a recent review of the physiological taste aversion literature (Bureš & Burešová, 1977). That review in large part covers the issues of concern here. Second, the physiological mechanisms of taste aversions are at this point far from clear. The studies performed most often train the taste aversions and perform the physiological manipulations in vastly different ways, and obtain vastly different results. For example, in studying hippocampal involvement in the acquisition of taste aversions, some researchers have lesioned the ventral or dorsal hippocampus, exposed the subjects once to a taste before pairing it with illness, and then assessed aversion using a one-bottle test (McGowan, Hankins, & Garcia, 1972), whereas others have aspirated most of the hippocampus, exposed the subjects seven times to a taste before pairing it with illness, and assessed aversion using a two-bottle test (Miller, Elkins, & Peacock, 1971). It is not surprising that the results of these studies were inconsistent, with Miller et al. but not McGowan et al. finding taste aversion disruption; reconciliation of their results is extremely difficult. Third, a difference in physiology between two learning tasks does not necessarily imply a difference in overt behavior. Even leaving aside discrepancies due to different sensory systems, activity in two separate areas of the brain can, perhaps through some later common neural path, result in the same external responses. It is those external responses that

we observe and monitor and by which the organism interacts with the rest of the world.

Associative Fluidity

Acquisition

Taste aversions are noted for forming in one trial. This unusual characteristic alone obliges their classification as prepared associations (Seligman, 1970). Testa and Ternes (1977) have suggested that one-trial taste aversion learning results from rats' lifetime experience with tastes and associated illnesses. Experiments with young organisms in which only one trial was needed for an aversion to form strongly suggest that Testa and Terne's explanation is not correct (Galef & Sherry, 1973; Grote & Brown, 1971; Rudy & Cheatle, 1977). There does indeed seem to be an inherited mechanism that enables taste–illness associations to be acquired rapidly.

Nevertheless, this prewiring for rapid learning may not be unique to the acquisition of taste aversions. One-trial learning can also occur, for example, in standard conditioning avoidance paradigms. Avoidance conditioning may occur rapidly, though only when the avoidance response, such as freezing, is the natural defense reaction of a species (Bolles, 1970). In that case the avoidance learning would also be classified as a prepared association.

Although the rapidity with which taste aversions are learned may not signify more than a parametric difference between taste aversion and other learning, it does make the process of acquisition more difficult to observe because of ceiling and floor effects (Kalat, 1977; Revusky & Garcia, 1970). Garcia et al. (1966) circumvented this constraint by use of a procedure that conditioned a weak aversion. They were thus able to observe the development of an aversion over several trials. The amount of saccharin consumed decreased as the number of saccharin–illness pairings increased in a way consistent with learning curves in more standard paradigms (Kimble, 1961; Mackintosh, 1974).

Retention

Once acquired, taste aversions are retained over an extremely long period of time. Experiments have shown retention for as long as 90 days (Dragoin, Hughes, Devine, & Bentley, 1973). Even day-old guinea pigs trained to avoid sucrose after one trial retained their aversions more than a month later (Kalat, 1975). But such long periods of retention are not unusual for aversively motivated learning in general. For example, Hoffman, Fleshler, and Jensen (1963) showed retention of conditioned suppression in pigeons to be virtually perfect 2.5 years after training. Gleitman (1971) extended these findings using rats, and he suggested that long retention may be the rule for classical aversive conditioning.

Extinction

To extinguish a conditioned response an experimenter presents the CS without the US. With repeated exposure to the unaccompanied CS, the conditioned response decreases in intensity or probability until it finally disappears. It has been suggested that taste aversions, since they are prepared, may be much more difficult to extinguish than traditional tasks (Mitchell, Scott, & Mitchell, 1977; Seligman, 1970; Seligman & Hager, 1972). Nevertheless, neither the overall process of extinction nor the rapidity with which it occurs appears discriminably different between taste aversion and traditional learning.

Numerous experiments have shown that extinction of a taste aversion will occur when the CS is repeatedly presented alone (e.g., Baum, Foidart, & Lapointe, 1974; Domjan, 1975; Garcia et al., 1966; Nowlis, 1974). In addition, it has been found that extinction is faster if exposure to the CS is facilitated by making the CS a subject's sole source of some substance of which the subject has been deprived. For example, Grote and Brown (1973) showed that with only one fluid source available water deprivation hastens extinction of a saccharin solution CS by increasing (safe) CS consumption during extinction trials. Similarly, aversion to a sodium chloride solution is extinguished faster when a need for sodium is produced by injections of formalin or by adrenalectomy, or when a need for water is induced by water deprivation or drug injections (Balagura & Smith, 1970). These experiments indicate that the degree of

extinction of taste aversions varies with the amount of CS-alone presentation, as does the degree of extinction in more traditional learning tasks (see also Abelson, Pierrel-Sorrentino, & Blough, 1977).

There is such a large range in the resistance to extinction of tasks acquired in standard conditioning procedures that to make a statement to the effect that taste aversions are more or less difficult to extinguish at best obscures the facts. Characteristics of experimental procedure during acquisition are influential in determining rapidity of extinction (for a review of the literature see Kimble, 1961; Mackintosh, 1974), and this influence also makes it difficult to compare taste aversion with traditional task extinction. A more fruitful approach may be to compare extinction of prepared and unprepared learning, equating as many aspects of the procedures for the two types of learning as possible. Some data relevant to such a comparison already exist (see Seligman & Hager, 1972). However, the ability of the researchers cited in the preceding paragraph to extinguish taste aversions without inordinate difficulties (in Garcia et al.'s, 1966, experiment extinction was complete after three 10-minute sessions) argues against a statement to the effect that taste aversions almost always take longer to extinguish than do responses acquired in traditional tasks.

Summary

Acquisition occurs over time as a CS and a US are presented together. Once CS–US pairings have ceased, responses to the CS gradually decrease, although complete cessation of responding may take years. This process occurs more quickly when the CS is presented without the US, that is, extinction occurs. This description applies equally well to both taste aversion and traditional learning. Further, the data point to no unique temporal aspects of these processes for taste aversion learning; taste aversions are not acquired faster, retained longer, or extinguished over more trials than all other traditional learning tasks. Previous studies of operant and classical conditioning have included a wide range of prepared and contraprepared associations. At most the research indicates that taste aversions are relatively easier to acquire, a quantitative, not a qualitative, difference.

CS and US Characteristics and Their Interaction

Intensity

US. With increasing US intensity, learned taste aversions are more pronounced (Dragoin, 1971; Nachman & Ashe, 1973; Revusky, 1968). This is consistent with other findings in the operant and classical conditioning literature. Ray and Bivens (1968) trained mice on a passive avoidance task and found that with a greater intensity US, learning was more persistent following amnesic electroconvulsive shock. In Passey's (1948) experiment with humans, conditioned eyelid responses were larger and were acquired more quickly when a stronger air puff was used as the US. Finally, Church (1969) has summarized evidence indicating that the extent to which punishment of an instrumental response suppresses that response is a positive function of the intensity of the punishment.

CS. The taste aversion CS also seems to support better conditioning when it is stronger (Barker, 1976; Dragoin, 1971; Nowlis, 1974). Pavlov (1927/1946) observed this phenomenon in his own experiments on salivary conditioning, and Hull (1949) incorporated it into his learning theories as stimulus intensity dynamism. Gray (1965) discussed the more recent evidence for this phenomenon in classical conditioning generally.

Specificity of CS to US

Taste aversion. Garcia and his colleagues, working with rats, were the first to report specificity of CS to US in taste aversion learning. The seminal study (Garcia & Koelling, 1966) found that if both an audiovisual and a taste cue are correlated with illness, only the taste readily becomes aversive. Similarly, if the two cues are paired with shock, only the audiovisual cue readily becomes aversive. Additional work (Garcia & Koelling, 1967; Garcia, McGowan, Ervin, & Koelling, 1968) showed that a taste was easier to

associate with illness than was an odor or the size of a piece of food. Green, Bouzas, and Rachlin (1972), using a more controlled procedure, have also demonstrated specificity of CS to US. They followed rats' saccharin drinking with 1 hour of poisoning, pulsed shock, or continuous shock. Only poisoning suppressed drinking.

The tendency of the taste of the food, rather than any other aspect, to become associated with illness is one of the best known characteristics of illness-induced learning. However, it should be remembered that despite occasional failures (Larsen & Hyde, 1977), it is possible for cues other than tastes to be associated with illness in rats, although often more trials and more careful training procedures are required (Riccio & Haroutunian, 1977; Rozin, 1969). P. J. Best, Best, and Henggeler (1977) summarized a good deal of evidence in support of this statement, but they also pointed out that illness-induced aversions to environmental cues are not as robust as taste aversions. The maximum delay of reinforcement is shorter, and the aversions extinguish faster.

Traditional learning. But what about more standard learning paradigms? Do they also show specificity of CS to US? The answer is yes, although examples are not as easy to find for them as they are for taste aversion learning. Shettleworth (1972b) used three different procedures in her experiments with chicks: punishment with foot shock for drinking water, punishment with shock in the water for drinking, or fear conditioning by pairing exteroceptive stimuli with foot shock while the chicks drank. For the first two procedures, a compound CS consisting of a flashing light plus clicking was associated with shock and controlled behavior. For the conditioned emotional response paradigm, the clicking controlled behavior. However, Shettleworth's studies were not ideal for showing CS to US specificity, as they used different responses and different experimental procedures.

Two operant conditioning experiments did not have this problem. Delius (1968) showed that when pigeons were thirsty and pecking for water reinforcement they were more likely to peck at stimulus cards toward the shortwave (blue) end of the spectrum relative to when they were hungry and pecking for food reinforcement. In Foree and LoLordo's (1973) experiments pigeons pressed a treadle either to obtain food or to avoid shock. Both visual and auditory discriminative stimuli were present. Visual stimuli were prepotent in controlling behavior when food was being obtained; auditory stimuli were prepotent when shock was being avoided.

What these results may indicate is not simply that standard learning can show instances of CS to US specificity, as occur in taste aversion acquisition, but that the frequency of prepared and contraprepared associations is greater and more pervasive than has been previously assumed. Clearly, the preparedness of associations can affect experimental results, and taste aversions are by no means the only example. Only additional experiments will determine the extent of CS to US specificity in traditional learning.

Differences between species. As a result of the influence of Darwin's theory of natural selection, all species are thought to have evolved from some small number of ancestors and thus to have many characteristics in common. This belief in the continuity of the species is fundamental for animal psychologists who spend their careers working with lower organisms, thereby developing principles that many of these researchers hope will one day be applied to human beings. Nevertheless, the taste aversion literature has pointed out some very striking species-specific differences. In particular, for some species gustatory cues do not serve as the most prepared CS in illness-induced learning. Instead, visual cues are prepotent. Wilcoxon, Dragoin, & Kral (1971) provided quail with blue sour water to drink and then poisoned the quail. Aversions were formed to the blue color and not to the sour taste of the fluid. This modality preference is adaptive, for quail's food consists of seeds that are covered by a hard, flavorless shell. To determine what is and is not poison, diurnal birds that consume seeds must use visual cues (Garcia et al., 1974; Gustavson, 1977).

It could be argued that the ability of quail to form visually based illness-induced aversions is solely attributable to the fact that the visual apparatus of quail, diurnal birds,

is so highly developed. But this argument cannot be made for results from taste aversion experiments that employed guinea pigs (Braveman, 1974, 1975b). This species has no difficulties associating visual cues with illness, despite the fact that their visual system is no more developed than that of the rat. In contrast, rats do not easily associate visual cues with illness. Guinea pigs and quail search for their food by day and use their visual systems in this search. Rats obtain food at night using their visual systems to a lesser extent. The stimuli of the modality used in searching for food may be those that are prepared to associate with illness rather than simply the stimuli perceived by the most highly developed sensory system (Braveman, 1974, 1975b).

Gustavson (1977) summarized much of the evidence on the formation of illness-based aversions in different species. He concluded that to a surprising extent, these aversions are formed in a similar manner across many species. The prepotent CS modality differs depending on the particular ecological niche of the species, but overall the effects of CS and US intensity, and the maximum interval between the CS and the US, are strikingly alike across species. It should also be noted that species differences in cue modality predominance are not confined to the taste aversion paradigm (see Kalat, 1977, for a discussion of this point).

CS salience as a function of the US. Not only are some CSs predisposed to be associated with some USs in taste aversion learning but, even more specifically, some CSs may be associated with injections of certain aversive substances and not others. Weisinger, Parker, and Skorupski (1974) have reported experiments showing that if saline is used as the CS it will become aversive if insulin, which decreases blood sugar, is used as the US. However, if formalin, which causes a need for sodium, is used as the US, an aversion to saline will not form. On the other hand, when sucrose is employed as the CS the opposite results are obtained. Domjan and Levy (1977) pointed out the possible evolutionary basis of this behavior: Animals low in blood sugar are unlikely to become ill while consuming sucrose, nor while consuming saline when they are low in salt. In keeping with these findings, Frumkin (1975) was unable to train adrenalectomized (adrenalectomy causes a decrease in body sodium) rats to avoid sodium chloride or to train parathyroidectomized (parathyroidectomy causes a decrease in body calcium) rats to avoid calcium lactate.

Contrary to Weisinger et al.'s and Frumkin's results, Domjan and Levy were able to obtain conditioning of sucrose and saline with both insulin and formalin. Domjan and Levy could not completely isolate the reason for the discrepancy in results, although they eliminated many possibilities. Trent and Kalat (1977) obtained data similar to those of Domjan and Levy. They trained a taste aversion to sodium chloride in sodium-deficient subjects. They attributed their success to the design of the experiment: During testing the rats did not have a strong sodium hunger that would have masked any aversion to sodium. Clearly, even tastes that possess some kind of a positive biological significance can be made aversive in a taste aversion paradigm. This is consistent with the finding that the vaginal, sexual attractant secretion of the female hamster is easily associated with illness in male hamsters (Johnston & Zahorik, 1975; Zahorik & Johnston, 1976).

Generalization of the CS

Variation along more than one stimulus attribute. Generalization of the CS does occur to substances that have tastes similar to the CS. Domjan (1975) reported that after saccharin had been made aversive through contingent poison, casein hydrolysate, but not vinegar, was also subsequently avoided. Domjan drew the plausible conclusion that casein hydrolysate tastes more like saccharin than does vinegar. In experiments in which rats actively ingested LiCl (Balagura & Smith, 1970; Nachman, 1963; D. F. Smith & Balagura, 1969), the taste of the LiCl became an aversive CS and the aversion generalized to NaCl, a substance that causes responses similar to LiCl in the chorda tympani nerve (Nachman, 1963). These results are comparable to those obtained in traditional

conditioning procedures (Sutherland & Mackintosh, 1971).

Variation along one stimulus attribute. By testing for the amount of aversion to stimuli similar to the CS except for variations in one CS attribute, it is possible to determine the shape of a taste aversion generalization gradient. Generalization curves found using the usual laboratory paradigms take one of two basic forms: a straight line with positive or negative slope (stimulus intensity dynamism) or a curve passing through a minimum at the CS. The former is obtained when generalization is investigated along an intensity continuum such as brightness or loudness. The latter is observed with continua that change qualitatively as they are varied, for example, wavelength and pitch.

Despite suggestions that taste aversion generalization curves may demonstrate no stimulus control (Seligman & Hager, 1972), similar curves, and therefore similar equations, describe results of generalization testing in taste aversion and traditional learning paradigms when comparable modalities are employed. Several investigators have performed illness-induced aversion generalization experiments by poisoning after consumption of one saturation of blue water (Czaplicki, Borrebach, & Wilcoxon, 1976), one saccharin concentration (Logue, Note 1), or one salt concentration (Nowlis, 1974) and then testing with varying saturations or concentrations. The results suggested stimulus intensity dynamism with greater aversion as the tested stimulus increased in strength (Czaplicki et al., 1976), a generalization curve that passed through a minimum at its center, showing less aversion to test stimuli either stronger or weaker than the CS (Logue, Note 1), or both (Nowlis, 1974). Since saccharin concentration was the only clearly nonintensity continuum used in these experiments (the taste of saccharin changes qualitatively as its concentration changes; Collier & Novell, 1967), the results are consistent with predictions from the traditional literature.

Hedonic Value of the CS

A much discussed question in psychology is whether a neutral stimulus paired with a reinforcer becomes a signal for that reinforcer or whether the previously neutral stimulus itself acquires some hedonic value (Gleitman, 1974). Taste aversion theorists have claimed that a taste aversion CS is unusual in that it changes in hedonic value when it is paired with the US (Garcia & Hankins, 1977; Garcia et al., 1974; Garcia, McGowan, & Green, 1972; Garcia, Rusiniak, & Brett, 1977). In taste aversion experiments, not only does the organism act as if the CS causes illness but the subject displays other behaviors toward the CS indicating the changed hedonic tone of the stimulus. Garcia, McGowan, and Green (1972) cited as evidence for this that animals avoid the CS from a taste aversion experiment no matter where they encounter it, whereas when a taste is paired with shock the taste is avoided only in the experimental chamber in which the conditioning took place. However, note that money, a traditional, generalized, conditioned reinforcer, is of value even in situations in which it has never been paired with a primary reinforcer.

Additional evidence for the changed hedonic value of a taste aversion CS comes from Garcia et al.'s (1977) report of their observations of predator–prey relationships. These researchers have studied predators' reactions to their natural prey after consumption of the prey has been paired with illness. When presented with a meat carcass previously associated with LiCl, besides showing retching movements, coyotes demonstrate what the authors call *conditioned disgust responses,* including urinating on, burying, or rolling on the meat. A third class of behaviors are *conflict responses,* abnormal responses to the prey that take the place of the usual prey–predator relationship.

These types of responses to the CS are not confined to the taste aversion paradigm. Hearst and Franklin (1977) showed that pigeons will withdraw from a response key that has a negative correlation with grain. Timberlake and Grant (1975) used the presence of a rat as a food predictor for another rat. Subjects in the experiment directed social behaviors, not eating responses, toward the predictive rat. The authors stated that

their experiment provides support for the hypothesis that in classical conditioning a whole set of related responses is conditioned, not just a single reflex. Further, the concept of changed hedonic value has been present for many years in a traditional theory of human motivation under the name of *functional autonomy* (Allport, 1937, chapter 7). According to Allport, some stimuli paired with reinforcement become reinforcing in their own right and reduce drives that are not necessarily the same as those reduced by the original primary reinforcer.

Summary

This section has made abundantly clear the problems that arise in trying to compare taste aversion with so-called traditional learning. Although it was possible to establish some general effects that occur for both taste aversions and traditional tasks, for example, better learning with more intense CSs and USs and generalization of the CS to similar stimuli, other areas were not as easy to evaluate. There is evidence that some target responses in standard learning paradigms are easier to acquire with particular reinforcers than are others. Similarly, there are suggestions in the literature that the CS at least sometimes elicits a complex of interrelated behaviors both for animals acquiring taste aversions and for animals acquiring responses in traditional tasks. Therefore, with respect to specificity of CS to US and the hedonic value of the CS, it is difficult to compare taste aversion learning with the whole of traditional learning. The function that the preceding sections serve is to point out that in these respects the acquisition of taste aversions is not unique.

Temporal Relationship of CS and US

Trace Conditioning

Taste aversions are usually conditioned by first presenting the CS and then after a delay ranging from minutes to hours, presenting the US. This is similar to the classical conditioning procedure of trace conditioning. In a taste aversion paradigm delays of about an hour are common laboratory procedure. Learning has even occurred under optimal conditions with delays of 24 hours (Etscorn & Stephens, 1973). In accordance with what has been reported in traditional learning experiments (Renner, 1964), it has consistently been found that the strength of a learned taste aversion is less with longer intervals between the CS and the US during acquisition (see, e.g., M. R. Best & Barker, 1977; Burešová & Bureš, 1974; Garcia et al., 1966; Kalat & Rozin, 1971; Nachman, 1970; Revusky, 1968; J. C. Smith & Roll, 1967; Wilcoxon, 1977).

Although most experiments in traditional learning have shown that conditioning is possible only with CS–US delays of at most a few seconds (Kimble, 1961), a few experiments have demonstrated learning with longer delays. The latter studies examined the delay of reinforcement gradient by employing long intertrial intervals in two ways: Either reinforcement on one trial determined whether a response would be reinforced on the next trial or reinforcement for a trial was simply delayed until the start of the next trial. The first type of study has been reported by Capaldi (1967), Petrinovich and Bolles (1957), Petrinovich, Bradford, and McGaugh (1965), Pschirrer (1972), and Tyler, Wortz, and Bitterman (1953). In Capaldi's study rats ran in a runway, and reinforcement was presented on alternating trials. Despite 24-hour intertrial intervals the rats learned to run slower on trials that would not receive reinforcement.

Lett (1973, 1974, 1975) performed the second type of long-delay, traditional learning experiments. For her studies she used T mazes and delays of up to 1 hour between when the rat made its choice and was removed from the maze and when it was put back in the maze for another trial and received its reinforcement for the first trial. In a recent investigation (Lett, 1977), reward was given to a subject in its home cage following a session in the T maze. Animals still learned with delays of up to 2 hours between removal from the T maze and reward in the home cage. Lett was able to obtain

long-delay learning only when she removed the animals from the apparatus for the delay.

Lett's experiments have been offered as support for Revusky's (1971, 1977a) concurrent interference theory of long-delay learning. Revusky has stated that the maximum delay possible between the CS and US is an inverse function of the number of additional associations to the CS and US, other than the reference CS–US association, that form during the delay. These additional associations interfere with CS–US learning. Taste aversions are learned over long delays because other stimuli predisposed to associate with illness infrequently intervene between the CS and US. On the other hand, in traditional conditioning procedures employing a light or tone as the CS, other visual and auditory stimuli are very likely to intervene between the CS and US unless the CS–US delay is very short (see Kalat & Rozin, 1971, and Rozin & Ree, 1972, for additional support and criticism of the interference theory of long-delay learning).

Theories other than Revusky's have been offered to explain the differences in maximum delay of reinforcement usually found between taste aversion and traditional learning. An explanation based on the presence of an aftertaste that would bring the CS closer to the US has been sufficiently laid to rest (Garcia, Hankins, Robinson, & Vogt, 1972; Garcia, McGowan, & Green, 1972; Revusky & Garcia, 1970; Rozin, 1969). An alternative explanation relies on evidence of the unusual duration of the gustatory memory trace (Krane & Wagner, 1975). Regardless of which theory one wishes to accept or propose, if there is a difference between taste aversion and standard conditioning with respect to maximum delay of reinforcement it is a quantitative one; the data in question involve simply longer or shorter delays with perhaps some degree of overlap between the two paradigms.

Backward Conditioning

Backward conditioning, learning when the US is presented just prior to the CS, was traditionally assumed to be impossible (Kimble, 1961). But recently some standard learning experiments using more sophisticated procedures have reported backward conditioning (Heth & Rescorla, 1973; Keith-Lucas & Guttman, 1975; Wagner & Terry, 1975). In these experiments, maximum delays between the US and the CS are in the order of seconds. Results from taste aversion experiments are strikingly different. Although the very first attempt to demonstrate backward conditioning was not successful (Garcia & Kimeldorf, 1957), since then backward conditioning has been repeatedly obtained. Still at issue, however, is the maximum delay between the US and CS that will still support learning. Values range from .5 to 12 hours (Barker, Suarez, & Gray, 1974; Boland, 1973; Domjan & Gregg, 1977; Scarborough, Whaley, & Rogers, 1964). Domjan and Gregg reported that the maximum delay depends on the intensity of the CS.

It is not clear why this discrepancy in maximum US–CS delays between taste aversion and more standard learning exists, but it may have something to do with US duration. The US–CS delays in the taste aversion experiments were calculated as the time between irradiation (or poison injection or intubation) and the opportunity to taste. This is an inaccurate estimate, since the US consists of the sensations of illness, not of the manipulation that causes the illness, and since the CS consists of the actual taste, not simply of the availability of a taste. Illness resulting from poison or irradiation may still have been present hours later when the CS was presented, or even hours after CS termination, for an effective combination of backward, simultaneous, and trace conditioning. The traditional experiments all used shock of brief duration. These hypotheses are difficult to test because of methodological problems in measuring illness and aftertastes (see Barker, Smith, & Suarez, 1977, for a more detailed discussion of these issues), but it should be noted that in general, as in traditional learning, the maximum delay with which acquisition will occur in a backward-conditioning taste aversion paradigm is shorter than in a trace-conditioning taste aversion paradigm.

Summary

The temporal relationships between the CS and the US that support learning are not qualitatively different for taste aversion and traditional learning. In both cases, whether the CS or the US occurs first, close temporal contiguity is more effective than very long delays, and trace conditioning is more effective than backward conditioning. Most often, though not always, taste aversion learning is supported by longer delays than are possible in traditional learning. Several attempts have been made to explain this difference by reliance on such factors as the amount of interfering associations during the delay or the duration of the US rather than by reliance on a unique process of taste aversion formation. Nevertheless, it is striking that taste aversions have been acquired in one trial with a 24-hour delay (Etscorn & Stephens, 1973); nothing even approaching this has been demonstrated in traditional paradigms. This distinction between taste aversion and traditional learning appears to be one instance of such a large quantitative difference that simply calling it a quantitative, and not a qualitative, difference seems inappropriate.

Obtaining and Processing Information

Novelty

CS. Taste aversion conditioning appears to be easier with certain tastes than with others. Kalat and Rozin (1970) used the term *salience* to describe this finding. However, as was discussed above, CS intensity affects CS conditionability, and Kalat and Rozin did not equate the subjective intensity of their solutions, clearly a difficult task. In a later experiment, Kalat (1974) found that the degree to which a solution was novel could account for both CS intensity and salience effects. Although in most prior experiments rats had no previous experience with the solutions presented to them, the solutions differed to a varying extent from what the rats were used to drinking, which was plain water. Kalat raised rats on water or a high concentration of a particular solute. He then allowed them to drink a low and a medium concentration of that solute followed by poison. Rats raised on water formed an aversion to the medium concentration, whereas rats raised on a high concentration formed an aversion to the low concentration. There are many additional experiments that have carefully demonstrated the difficulty in associating a taste with illness if the taste has previously been presented without illness (e.g., M. R. Best & Gemberling, 1977; Fenwick, Mikulka, & Klein, 1975; Kalat & Rozin, 1973; Kiefer, Phillips, & Braun, 1977; Revusky & Bedarf, 1967; Siegel, 1974). This is comparable to the procedure in classical conditioning known as latent inhibition. There too a CS presented without a US, and before any presentation of a US, is more difficult to condition with that US (see Lubow, 1973, for a review of the literature).

There has recently been a great deal of controversy in the taste aversion literature over whether noncontingent illness can by itself cause an aversion, or an increased aversion, to novel foods similar to the process of sensitization in classical conditioning (Bitterman, 1976; Garcia, Hankins, & Rusiniak, 1976; Mitchell, 1977; Mitchell et al., 1977; Revusky, 1977b). Appeals to sensitization effects have been used in attempts to explain away the unusual associative properties of tastes with illness. Current evidence indicates that there are two types of enhanced aversion to novel foods following illness (Domjan, 1977). The first is a short-lived aversion to all novel foods, as a direct effect of the poison-induced illness. This phenomenon is observed after taste-contingent or taste-noncontingent poisoning and thus is comparable to sensitization. Second, after taste-contingent poison presentation, a more lasting aversion occurs to novel foods that taste similar to the taste paired with illness. The aversion to the taste paired with illness generalizes to similar novel foods, an associative process (i.e., not sensitization). Therefore if subjects are completely recovered from illness before any aversion testing begins, it is extremely unlikely that any sensitization effects will be observed. Most ex-

perimenters have indeed waited until the animals were no longer ill before beginning testing (for similar opinions and supporting evidence, see Garcia & Hankins, 1977; Revusky, 1977c; Rozin, 1977; Testa & Ternes, 1977; Wilcoxon, 1977; Zahorik, 1977).

US. Parallel to the finding with CSs and contrary to what sensitization would predict, if a US is presented alone and not correlated with a CS, subsequent conditioning of a taste aversion with that US is retarded (Cannon, Berman, Baker, & Atkinson, 1975; Cappell, LeBlanc, & Herling, 1975; Elkins, 1974; Kiefer et al., 1977; Mikulka, Leard, & Klein, 1977). This was demonstrated even when tolerance and addiction effects were controlled for (Bravemen, 1975a; Vogel, Note 2). The complementary US preexposure effect has been noted in traditional conditioning experiments (e.g., Ayres, Benedict, & Witcher, 1975; Mis & Moore, 1973; Siegel & Domjan, 1971).

Learned Irrelevance

Learned irrelevance is the term that Mackintosh (1973, 1974) used to describe an organism's learning that a given CS and US are not correlated. This occurs with random presentations of both the CS and the US and is more deleterious to further conditioning than simply presenting the CS alone (latent inhibition). Mackintosh cited a great deal of evidence to show that organisms can learn a lack of correlation between the CS and the US. So far this phenomenon has not been explicitly investigated in taste aversion learning. This is probably due to the problems inherent in randomly presenting CSs and USs without having some association take place between them as a result of the long delays of reinforcement possible with taste aversions. That taste aversions do not form with extremely long CS–US delays, presumably because the organism learns that the CS is not associated with poison, has been described as demonstrating learned irrelevance in a taste aversion paradigm (Kalat, 1977).

Blocking

There are at least two basic procedures with which to study the blocking effect. Kamin originally reported blocking in 1969 upon showing that prior conditioning with one member of a compound stimulus blocked later conditioning to the other member of the compound when the compound was paired with a US. Revusky (1971) obtained blocking in a taste aversion experiment using this type of a blocking paradigm and taste stimuli. Kalat and Rozin (1972), also employing taste stimuli, were unable to confirm this result; the novel member of the compound stimulus paired with the US still became aversive. However, subsequent experiments have demonstrated that this type of blocking does occur in taste aversion learning (Gillan & Domjan, 1977; Rudy, Iwens, & Best, 1977). Rudy et al. first paired novel exteroceptive cues with illness. Later, acquisition of an aversion to a taste was impeded when the compound of the taste and the familiar exteroceptive stimulation were paired with illness. Gillan and Domjan used only oral stimuli in their experiments.

The second type of procedure used in demonstrating blocking was employed by Wagner, Logan, Haberlandt, and Price (1968). In this method one compound containing two stimuli is paired with the US, and another compound containing one stimulus from the first compound plus a third stimulus is presented but not paired with the US. A test for learning is then conducted with the stimulus common to both compounds. Subjects in Wagner et al.'s experiments demonstrated blocking by responding less to the common stimulus under these reinforcement conditions than when each compound was followed by the US half the time. Luongo (1976) obtained similar results from a taste aversion experiment.

There have been other attempts to manipulate the amount of information provided by a taste CS paired with an illness US (see Kalat & Rozin, 1972; Revusky, Parker, & Coombes, 1977), but since comparable traditional designs have yet to be performed, it is difficult to interpret the taste aversion results.

Conditioned Inhibition

A conditioned inhibitor is a cue that guarantees that reinforcement will not occur when it might otherwise be expected to occur; it has a negative correlation with reinforcement (Kimble, 1961; Mackintosh, 1974). Conditioned inhibition has been demonstrated in taste aversion. M. R. Best (1975) paired saccharin with illness and then saline with no illness. The subjects subsequently chose between saline and some third liquid. Their preference for saline as compared with the other liquid available was greater than that of control animals. In Taukulis and Revusky's (1975) experiment saccharin was always followed by illness, but saccharin accompanied by an odor was never followed by illness. The odor then satisfied three of Rescorla's (1969, 1971) conditions for classification as a conditioned inhibitor: The presentation of the odor with a CS inhibited the conditioned response to the CS, aversion conditioning of the odor was harder than aversion conditioning of a neutral stimulus, and the odor enhanced conditioning to a neutral stimulus when odor and neutral stimulus presented together were followed by reinforcement.

Sensory Preconditioning

The first step in attempting sensory preconditioning is to pair two stimuli. One of them is subsequently reinforced. If the other, nonreinforced stimulus also acquires motivational properties, as if it too had been paired with the reinforcer, sensory preconditioning has taken place. This phenomenon has frequently been demonstrated in traditional laboratory experiments (see, e.g., Brogden, 1939; Rizley & Rescorla, 1972). It has also been shown to occur in a taste aversion paradigm (Lavin, 1976). Lavin exposed rats to two tastes, one right after the other. He then followed a presentation of one of these tastes with illness. The taste that was not paired with illness became aversive. In addition, this example of sensory preconditioning in a taste aversion paradigm occurred only if during the original pairing the two tastes were separated by no more than a few seconds, as in traditional paradigms.

Second-Order Conditioning

Second-order conditioning is similar to sensory preconditioning except that the two stimuli are presented together after one of them has been paired with illness instead of before the pairing (see Rizley & Rescorla, 1972). Although it has been difficult to establish the effect in traditional classical conditioning experiments, Rescorla's (1977) recent experiments have been successful. Second-order conditioning has also been shown for taste aversion learning. P. J. Best, Best, and Mickley (1973) paired a taste cue with a visual cue previously made aversive through contingent illness. The taste was subsequently avoided. Bond and Harland (1975) used a similar procedure, but employed only tastes as stimuli; they also obtained second-order conditioning.

Summary

The ways in which information is obtained and processed could potentially have revealed qualitative as well as quantitative differences between taste aversion and traditional learning. For example, had blocking, learned irrelevance, and conditioned inhibition not been found to occur within taste aversion paradigms, a good case could have been made for the necessity of using different basic principles to describe the learning of taste aversions. Nevertheless, the data appear to show that in every respect the acquisition and manipulation of information are quite similar for taste aversion and traditional learning.

Age Effects

Several experimenters who used a taste aversion paradigm have reported learning in rat pups (Ader & Peck, 1977; Galef & Sherry, 1973; Grote & Brown, 1971; Klein, Mikulka, Domato, & Hallstead, 1977; Rudy & Cheatle, 1977). One trial is often all that is necessary for learning to occur (Ader & Peck, 1977; Galef & Sherry, 1973; Grote & Brown, 1971;

Rudy & Cheatle, 1977). This seems inconsistent with results obtained in traditional learning in which acquisition of passive avoidance tasks, a paradigm similar in many ways to the taste aversion paradigm, is worse in weanling than in adult rats (Brunner, 1969; Campbell & Coulter, 1976; Riccio, Rohrbaugh, & Hodges, 1968). However, it would be premature to conclude, based on this evidence alone, that weanling rats learn taste aversions as easily as adults, but learn passive avoidance tasks much worse than adults.

Although young rats do learn taste aversions, it may not be as easy for them to acquire aversions as it is for adult rats (Klein, Domato, Hallstead, Stephens, & Mikulka, 1975). On the traditional learning paradigm side of the question, Feigley and Spear (1970) reported an experiment in which they showed that weanling rats do as well as adults at a passive avoidance task if the young rats are trained in an apparatus proportional to their body size. The question of the comparative conditionability of weanling and adult rats in standard and taste aversion learning paradigms is far from settled.

Similarly, it first appeared that rat pups, unlike adults, show little long-term retention of all types of tasks (Campbell & Coulter, 1976; Feigley & Spear, 1970) except taste aversions (Ader & Peck, 1977; Klein et al., 1977). However, a recent article by Coulter, Collier, and Campbell (1976) has shown that if a conditioned emotional response paradigm is used in which the animal's size is irrelevant to the performance of the task (as may be the case with taste aversion learning), long-term retention is found even when the animals are trained at a much earlier age (17 days) than was previously thought possible for traditional conditioning. The rats in Coulter et al.'s studies were at least as young as those used in the taste aversion retention experiments.

Conclusions

A detailed comparison of taste aversion data with results from more standard paradigms is significant not for the number of differences revealed between the two research areas, but for the number of similarities. In virtually all cases the same principles are sufficient for describing taste aversion and traditional learning data. In addition to the qualitative similarities, the two research areas are also often quantitatively similar. To take a simple example, taste aversions are extinguished by presenting the CS without the US, as in a standard classical conditioning experiment; similar principles apply for both paradigms. Further, several and sometimes many trials are needed for extinction of a taste aversion to occur, numbers that apply as readily to tasks learned in traditional paradigms.

Some qualifications are in order, however. Two problems of comparing taste aversion with traditional learning were stated in the Introduction. These were, first, defining the laws of learning and, second, specifying what constituted a deviation from one of these laws. After reviewing the literature, a third difficulty became apparent. The research discussed in the section entitled Traditional Learning is not homogeneous. Within that category are many examples of what Seligman (1970; Seligman & Hager, 1972) would call prepared and contraprepared learning. In addition, much of traditional learning also involves feeding or interoceptive stimuli. The characteristics of taste aversion learning that are said to be unique appear frequently to be better described as characteristics of prepared or feeding behavior. Any comparison of taste aversion and traditional learning should take these factors into account. Nevertheless, this review has demonstrated that at the level of the whole organism there is little basis at this time for a claim that there are qualitative differences between taste aversion learning and traditional learning, as best demonstrated by the sections entitled Obtaining and Processing Information and Generalization of the CS.

At several points in this article the data suggested that taste aversion learning shows certain quantitative differences from the learning of most other traditional tasks. Included among these are differences concerning rapidity of acquisition and long-delay

learning. Although one might say that these are only parametric peculiarities, they are important determinants of behavior, which may be evolutionarily determined, and they should not be ignored. It was also noted that the fact that taste aversions are acquired in one trial with extremely long delays is such a large parametric difference that in this instance the qualitative–quantitative distinction appears inappropriate.

To be able to describe and predict what occurs in taste aversion learning, it is necessary to have a full understanding of each species' feeding behavior under natural conditions. Otherwise, one would be unlikely to predict that, for example, rats will more easily associate tastes than exteroceptive stimuli with illness. The unusual quantitative properties of taste aversion learning mentioned above are in accord with what is adaptive for the organism (Kalat, 1977). Credit must be given to those who are calling for more examination of laws of learning and behavior with regard to these laws' adaptiveness for a species' ecological niche (e.g., Bolles, 1973; Garcia et al., 1977; Hinde, 1973; Lockard, 1971; Rozin, 1976, 1977; Schwartz, 1974; Seligman & Hager, 1972).

Many of these authors have in addition disparaged some of the more traditional learning theorists, who they claim stated that any stimulus is equally associable with any response and that laws of learning are the same across all species. However, the traditional theorists very often did realize that there are differences between species, but thought that the similarities between species are the more interesting and more important phenomena to study in developing a science of behavior (see, e.g., Skinner, 1959, pp. 374–375). Other psychologists unfortunately too often interpreted this belief as a theoretical assumption of more similarity than the data ever indicated.

Recent emphasis on discoveries of diversity and complexity in the characteristics of learning does not mean that an attempt must now be made to describe new principles and laws. Instead, in at least some cases the traditional ones should be made flexible enough to deal successfully with the species- and task-specific characteristics that are more prevalent than previously recognized (see Krane & Wagner, 1975; Rozin, 1977; and Seligman, 1973, for similar views). Without a system of general laws we might find ourselves in an overwhelming jungle of unique abilities and behaviors. In addition, a hypothesis of general laws of learning is not necessarily unreasonable from an evolutionary standpoint. Since every species has been shaped by evolution and since certain selection pressures are common to us all, species-general behaviors could result (Lockard, 1971; Rozin & Kalat, 1971). Further, phyletic closeness (Lockard, 1971), economy of neural wiring of the organism (Rozin & Kalat, 1971), or the uniformity of rules of future prediction (Revusky, 1977d) may give rise to similar processes of learning in different species and in different situations.

The effects that taste aversion experiments have had on preexisting psychological theory are consistent with Kuhn's (1962) description of the progress of scientific revolutions. Anomalous taste aversion data challenged traditional theory (Rozin, 1977). These data were different enough to impel some psychologists (e.g., Seligman, 1970) to suggest that many of the standard laws would not apply. Others (e.g., Rozin & Kalat, 1971) advocated the replacement of the old paradigm, general laws of learning, with a new one, a principle of evolutionary adaptiveness. A thorough comparison of the taste aversion and traditional learning data now reveals them not to be so discrepant after all. Yet without the emphasis by various learning theorists on the anomalous character of the taste aversion results, it is doubtful whether the somewhat unusual aspects of these data would ever have been noticed (see also Boring, 1929).

After the proposal of general laws of learning and a recent swing to laws that are unique and specialized, a less extreme swing, back to the eclectic middle ground, is due. This view recognizes the existence and utility of general laws of learning, but it also recognizes the necessity of acknowledging and investigating the dissimilarities in the learning of different species and in the learning of dif-

ferent tasks. Otherwise we are likely to assume generality where none exists. The research on taste aversion has taught us at least this much.

Reference Notes

1. Logue, A. W. *Generalization of the conditioned stimulus in taste aversion learning*. Paper presented at the meeting of the Eastern Psychological Association, Washington, D.C., April 1978.
2. Vogel, J. R. *Prior exposure to a drug (US) attenuates learned taste aversions*. Paper presented at the meeting of the Psychonomic Society, Boston, November 1974.

References

Abelson, J. S., Pierrel-Sorrentino, R., & Blough, P. M. Some conditions for the rapid extinction of a learned taste aversion. *Bulletin of the Psychonomic Society*, 1977, *9*, 51–52.

Ader, R., & Peck, J. H. Early learning and retention of a conditioned taste aversion. *Developmental Psychobiology*, 1977, *10*, 213–218.

Allport, G. W. *Personality: A psychological interpretation*. New York: Holt, 1937.

Ayres, J. J. B., Benedict, J. O., & Witcher, E. S. Systematic manipulation of individual events in a truly random control in rats. *Journal of Comparative and Physiological Psychology*, 1975, *88*, 97–103.

Balagura, S., & Smith, D. F. Role of LiCl and environmental stimuli on generalized learned aversion to NaCl in the rat. *American Journal of Physiology*, 1970, *219*, 1231–1234.

Barker, L. M. CS duration, amount, and concentration effects in conditioning taste aversions. *Learning and Motivation*, 1976, *7*, 265–273.

Barker, L. M., Smith, J. C., & Suarez, E. M. "Sickness" and the backward conditioning of taste aversions. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977.

Barker, L. M., Suarez, E. M., & Gray, D. Backward conditioning of taste aversions in rats using cyclophosphamide as the US. *Physiological Psychology*, 1974, *2*, 117–119.

Barnett, S. A. *The rat: A study in behavior*. Chicago: University of Chicago Press, 1975.

Baum, M., Foidart, D. S., & Lapointe, A. Rapid extinction of a conditioned taste aversion following unreinforced intraperitoneal injection of the fluid CS. *Physiology & Behavior*, 1974, *12*, 871–873.

Best, M. R. Conditioned and latent inhibition in taste-aversion learning: Clarifying the role of learned safety. *Journal of Experimental Psychology: Animal Behavior Processes*, 1975, *1*, 97–113.

Best, M. R., & Barker, L. M. The nature of "learned safety" and its role in the delay of reinforcement gradient. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977.

Best, M. R., & Gemberling, G. A. Role of short-term processes in the conditioned stimulus preexposure effect and the delay of reinforcement gradient in long-delay taste-aversion learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 1977, *3*, 253–263.

Best, P. J., Best, M. R., & Henggeler, S. The contribution of environmental non-ingestive cues in conditioning with aversive internal consequences. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977.

Best, P. J., Best, M. R., & Mickley, G. A. Conditioned aversion to distinct environmental stimuli resulting from gastrointestinal distress. *Journal of Comparative and Physiological Psychology*, 1973, *85*, 250–257.

Bitterman, M. E. The comparative analysis of learning. *Science*, 1975, *188*, 699–709.

Bitterman, M. E. Flavor aversion studies. *Science*, 1976, *192*, 266–267.

Boland, F. J. Saccharine aversions induced by lithium chloride toxicosis in a backward conditioning paradigm. *Animal Learning & Behavior*, 1973, *1*, 3–4.

Bolles, R. C. Species-specific defense reactions and avoidance learning. *Psychological Review*, 1970, *77*, 32–48.

Bolles, R. C. The comparative psychology of learning: The selective association principle and some problems with "general" laws of learning. In G. Bermant (Ed.), *Perspectives on animal behavior*. Glenview, Ill.: Scott, Foresman, 1973.

Bond, N., & Harland, W. Higher order conditioning of a taste aversion. *Animal Learning & Behavior*, 1975, *3*, 295–296.

Boring, E. G. The psychology of controversy. *Psychological Review*, 1929, *36*, 97–121.

Braveman, N. S. Poison-based avoidance learning with flavored or colored water in guinea pigs. *Learning and Motivation*, 1974, *5*, 182–194.

Braveman, N. S. Formation of taste aversions in rats following prior exposure to sickness. *Learning and Motivation*, 1975, *6*, 512–534. (a)

Braveman, N. S. Relative salience of gustatory and visual cues in the formation of poison-based food aversions by guinea pigs (*Cavia porcellus*). *Behavioral Biology*, 1975, *14*, 189–199. (b)

Breland, K., & Breland, M. The misbehavior of organisms. *American Psychologist*, 1961, *16*, 681–684.

Brogden, W. J. Sensory pre-conditioning. *Journal of Experimental Psychology*, 1939, *25*, 323–332.

Brunner, R. L. Age differences in one-trial passive avoidance learning. *Psychonomic Science*, 1969, *14*, 134.

Bureš, J., & Burešová, O. Physiological mechanisms

of conditioned food aversion. In N. W. Milgram, L. Krames, & T. M. Alloway (Eds.), *Food aversion learning*. New York: Plenum Press, 1977.

Burešová, O., & Bureš, J. Functional decortication in the CS–US interval decrease efficiency of taste aversive learning. *Behavioral Biology*, 1974, *12*, 357–364.

Campbell, B. A., & Coulter, X. Neural and psychological processes underlying the development of learning and memory. In T. J. Tighe & R. N. Leaton (Eds.), *Habituation*. Hillsdale, N. J.: Erlbaum, 1976.

Cannon, D. S., Berman, R. F., Baker, T. B., & Atkinson, C. A. Effect of preconditioning unconditioned stimulus experience on learned taste aversions. *Journal of Experimental Psychology: Animal Behavior Processes*, 1975, *1*, 270–284.

Capaldi, E. J. A sequential hypothesis of instrumental learning. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 1). New York: Academic Press, 1967.

Cappell, H., LeBlanc, A. E., & Herling, S. Modification of the punishing effects of psychoactive drugs in rats by previous drug experience. *Journal of Comparative and Physiological Psychology*, 1975, *89*, 347–356.

Capretta, P. J. An experimental modification of food preference in chickens. *Journal of Comparative and Physiological Psychology*, 1961, *54*, 238–242.

Church, R. M. Response suppression. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior*. New York: Appleton-Century-Crofts, 1969.

Collier, G., & Novell, K. Saccharin as a sugar surrogate. *Journal of Comparative and Physiological Psychology*, 1967, *64*, 404–408.

Coulter, X., Collier, A. C., & Campbell, B. A. Long-term retention of early Pavlovian fear conditioning in infant rats. *Journal of Experimental Psychology: Animal Behavior Processes*, 1976, *2*, 48–56.

Czaplicki, J. A., Borrebach, D. E., & Wilcoxon, H. C. Stimulus generalization of an illness-induced aversion to different intensities of colored water in Japanese quail. *Animal Learning & Behavior*, 1976, *4*, 45–48.

Delius, J. D. Color preference shift in hungry and thirsty pigeons. *Psychonomic Science*, 1968, *13*, 273–274.

Domjan, M. Poison-induced neophobia in rats: Role of stimulus generalization of conditioned taste aversions. *Animal Learning & Behavior*, 1975, *3*, 205–211.

Domjan, M. Attenuation and enhancement of neophobia for edible substances. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977.

Domjan, M., & Gregg, B. Long-delay backward taste-aversion conditioning with lithium. *Physiology & Behavior*, 1977, *18*, 59–62.

Domjan, M., & Levy, C. J. Taste aversions conditioned by the aversiveness of insulin and formalin: Role of CS specificity. *Journal of Experimental Psychology: Animal Behavior Processes*, 1977, *3*, 119–131.

Dragoin, W. B. Conditioning and extinction of taste aversions with variations in intensity of the CS and UCS in two strains of rats. *Psychonomic Science*, 1971, *22*, 303–305.

Dragoin, W. B., Hughes, G., Devine, M., & Bentley, J. Long-term retention of conditioned taste aversions: Effects of gustatory interference. *Psychological Reports*, 1973, *33*, 511–514.

Elkins, R. L. Bait-shyness acquisition and resistance to extinction as functions of US exposure prior to conditioning. *Physiological Psychology*, 1974, *2*, 341–343.

Etscorn, F., & Stephens, R. Establishment of conditioned taste aversions with a 24-hour CS-US interval. *Physiological Psychology*, 1973, *1*, 251–253.

Feigley, D. A., & Spear, N. E. Effects of age and punishment condition on long-term retention by the rat of active- and passive-avoidance learning. *Journal of Comparative and Physiological Psychology*, 1970, *73*, 515–526.

Fenwick, S., Mikulka, P. J., & Klein, S. B. The effect of different levels of pre-exposure to sucrose on the acquisition of a conditioned aversion. *Behavioral Biology*, 1975, *14*, 231–235.

Foree, D. D., & LoLordo, V. M. Attention in the pigeon: Differential effects of food-getting versus shock-avoidance procedures. *Journal of Comparative and Physiological Psychology*, 1973, *85*, 551–558.

Frumkin, K. Failure of sodium- and calcium-deficient rats to acquire conditioned taste aversions to the object of their specific hunger. *Journal of Comparative and Physiological Psychology*, 1975, *89*, 329–339.

Galef, B. G., & Sherry, D. F. Mother's milk: A medium for transmission of cues reflecting the flavor of mother's diet. *Journal of Comparative and Physiological Psychology*, 1973, *83*, 374–378.

Garcia, J., Ervin, F. R., & Koelling, R. A. Learning with prolonged delay of reinforcement. *Psychonomic Science*, 1966, *5*, 121–122.

Garcia, J., & Hankins, W. G. On the origin of food aversion paradigms. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977.

Garcia, J., Hankins, W. G., Robinson, J. H., & Vogt, J. L. Bait shyness: Tests of CS–US mediation. *Physiology & Behavior*, 1972, *8*, 807–810.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. Behavioral regulation of the milieu interne in man and rat. *Science*, 1974, *185*, 824–831.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. Flavor aversion studies. *Science*, 1976, *192*, 265–266.

Garcia, J., & Kimeldorf, D. J. Temporal relationship

within the conditioning of a saccharine aversion through radiation exposure. *Journal of Comparative and Physiological Psychology,* 1957, *50,* 180–183.

Garcia, J., Kimeldorf, D. J., & Hunt, E. L. The use of ionizing radiation as a motivating stimulus. *Psychological Review,* 1961, *68,* 383–395.

Garcia, J., Kimeldorf, D. J., & Koelling, R. A. Conditioned aversion to saccharin resulting from exposure to gamma radiation. *Science,* 1955, *122,* 157–158.

Garcia, J., & Koelling, R. A. Relation of cue to consequence in avoidance learning. *Psychonomic Science,* 1966, *4,* 123–124.

Garcia, J., & Koelling, R. A comparison of aversions induced by x-rays, toxins, and drugs in the rat. *Radiation Research Supplement,* 1967, 7 439–450.

Garcia, J., McGowan, B. K., Ervin, F. R., & Koelling, R. A. Cues: Their relative effectiveness as a function of the reinforcer. *Science,* 1968, *160,* 794–795.

Garcia, J., McGowan, B. K., & Green, K. F. Biological constraints on conditioning. In M. E. P. Seligman & J. L. Hager (Eds.), *Biological boundaries of learning.* New York: Appleton-Century-Crofts, 1972.

Garcia, J., Rusiniak, K. W., & Brett, L. P. Conditioning food-illness aversions in wild animals: *Caveant canonici.* In H. Davis & H. M. B. Hurwitz (Eds.), *Operant-Pavlovian interactions.* Hillsdale, N.J.: Erlbaum, 1977.

Gillan, D. J., & Domjan, M. Taste-aversion conditioning with expected versus unexpected drug treatment. *Journal of Experimental Psychology: Animal Behavior Processes,* 1977, *3,* 297–309.

Gleitman, H. Forgetting of long-term memories in animals. In W. K. Honig & P. H. R. James (Eds.), *Animal memory.* New York: Academic Press, 1971.

Gleitman, H. Getting animals to understand the experimenter's instructions. *Animal Learning & Behavior,* 1974, *2,* 1–5.

Gray, J. A. Stimulus intensity dynamism. *Psychological Bulletin,* 1965, *63,* 180–196.

Green, L., Bouzas, A., & Rachlin, H. Test of an electric-shock analog to illness-induced aversion. *Behavioral Biology,* 1972, *7,* 513–518.

Grote, F. W., & Brown, R. T. Rapid learning of passive avoidance by weanling rats: Conditioned taste aversion. *Psychonomic Science,* 1971, *25,* 163–164.

Grote, F. W., & Brown, R. T. Deprivation level affects extinction of a conditioned taste aversion. *Learning and Motivation,* 1973, *4,* 314–319.

Gustavson, C. R. Comparative and field aspects of learned food aversions. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection.* Waco, Tex.: Baylor University Press, 1977.

Hearst, E., & Franklin, S. R. Positive and negative relations between a signal and food: Approach-withdrawal behavior to the signal. *Journal of Experimental Psychology: Animal Behavior Processes,* 1977, *3,* 37–52.

Herrnstein, R. J. The evolution of behaviorism. *American Psychologist,* 1977, *32,* 593–603.

Heth, C. D., & Rescorla, R. A. Simultaneous and backward fear conditioning in the rat. *Journal of Comparative and Physiological Psychology,* 1973, *82,* 434–443.

Hinde, R. A. Constraints on learning: An introduction to the problems. In R. A. Hinde & J. Stevenson-Hinde (Eds.), *Constraints on learning.* New York: Academic Press, 1973.

Hoffman, H. S., Fleshler, M., & Jensen, P. Stimulus aspects of aversive controls: The retention of conditioned suppression. *Journal of the Experimental Analysis of Behavior,* 1963, *6,* 575–583.

Hull, C. L. The place of innate individual and species differences in a natural-science theory of behavior. *Psychological Review,* 1945, *52,* 55–60.

Hull, C. L. Stimulus intensity dynamism (*V*) and stimulus generalization. *Psychological Review,* 1949, *56,* 67–76.

Johnston, R. E., & Zahorik, D. M. Taste aversions to sexual attractants. *Science,* 1975, *189,* 893–894.

Kalat, J. W. Taste salience depends on novelty, not concentration, in taste-aversion learning in the rat. *Journal of Comparative and Physiological Psychology,* 1974, *86,* 47–50.

Kalat, J. W. Taste-aversion learning in infant guinea pigs. *Developmental Psychobiology,* 1975, *8,* 383–387.

Kalat, J. W. Biological significance of food aversion learning. In N. W. Milgram, L. Krames, & T. M. Alloway (Eds.), *Food aversion learning.* New York: Plenum Press, 1977.

Kalat, J. W., & Rozin, P. "Salience:" A factor which can override temporal contiguity in taste-aversion learning. *Journal of Comparative and Physiological Psychology,* 1970, *71,* 192–197.

Kalat, J. W., & Rozin, P. Role of interference in taste-aversion learning. *Journal of Comparative and Physiological Psychology,* 1971, *77,* 53–58.

Kalat, J. W., & Rozin, P. You can lead a rat to poison but you can't make him think. In M. E. P. Seligman & J. L. Hager (Eds.), *Biological boundaries of learning.* New York: Appleton-Century-Crofts, 1972.

Kalat, J. W., & Rozin, P. "Learned safety" as a mechanism in long-delay learning in rats. *Journal of Comparative and Physiological Psychology,* 1973, *83,* 198–207.

Kamin, L. J. Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior.* New York: Appleton-Century-Crofts, 1969.

Keith-Lucas, T., & Guttman, N. Robust-single-trial delayed backward conditioning. *Journal of Comparative and Physiological Psychology,* 1975, *88,* 468–476.

Kiefer, S. W., Phillips, J. A., & Braun, J. J. Preexposure to conditioned and unconditioned stim-

uli in taste aversion learning. *Bulletin of the Psychonomic Society*, 1977, *10*, 226–228.

Kimble, G. A. *Hilgard and Marquis' conditioning and learning*. New York: Appleton-Century-Crofts, 1961.

Klein, S. B., Domato, G. C., Hallstead, C., Stephens, I., & Mikulka, P. J. Acquisition of a conditioned aversion as a function of age and measurement technique. *Physiological Psychology*, 1975, *3*, 379–384.

Klein, S. B., Mikulka, P. J., Domato, G. C., & Hallstead, C. Retention of internal experiences in juvenile and adult rats. *Physiological Psychology*, 1977, *5*, 63–66.

Krane, R. V., & Wagner, A. R. Taste aversion learning with a delayed shock US: Implications for the "generality of the laws of learning." *Journal of Comparative and Physiological Psychology*, 1975, *88*, 882–889.

Kuhn, T. S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1962.

Larsen, J. D., & Hyde, T. S. A comparison of learned aversions to gustatory and exteroceptive cues in rats. *Animal Learning & Behavior*, 1977, *5*, 17–20.

Lavin, M. J. The establishment of flavor-flavor associations using a sensory preconditioning training procedure. *Learning and Motivation*, 1976, *7*, 173–183.

Lett, B. T. Delayed reward learning: Disproof of the traditional theory. *Learning and Motivation*, 1973, *4*, 237–246.

Lett, B. T. Visual discrimination learning with a 1-minute delay of reward. *Learning and Motivation*, 1974, *5*, 174–181.

Lett, B. T. Long delay learning in the T-maze. *Learning and Motivation*, 1975, *6*, 80–90.

Lett, B. T. Long delay learning in the T-maze: Effect of reward given in the home cage. *Bulletin of the Psychonomic Society*, 1977, *10*, 211–214.

Lockard, R. B. Reflections on the fall of comparative psychology: Is there a message for us all? *American Psychologist*, 1971, *26*, 168–179.

Lubow, R. E. Latent inhibition. *Psychological Bulletin*, 1973, *79*, 398–407.

Luongo, A. F. Stimulus selection in discriminative taste-aversion learning in the rat. *Animal Learning & Behavior*, 1976, *4*, 225–230.

Mackintosh, N. J. Stimulus selection: Learning to ignore stimuli that predict no change in reinforcement. In R. A. Hinde & J. Stevenson-Hinde (Eds.), *Constraints on learning*. New York: Academic Press, 1973.

Mackintosh, N. J. *The psychology of animal learning*. New York: Academic Press, 1974.

McGowan, B. K., Hankins, W. G., & Garcia, J. Limbic lesions and control of the internal and external environment. *Behavioral Biology*, 1972, *7*, 841–852.

Mikulka, P. J., Leard, B., & Klein, S. B. Illness-alone exposure as a source of interference with the acquisition and retention of a taste aversion. *Journal of Experimental Psychology: Animal Behavior Processes*, 1977, *3*, 189–201.

Miller, C. R., Elkins, R. L., & Peacock, L. J. Disruption of a radiation-induced preference shift by hippocampal lesions. *Physiology & Behavior*, 1971, *6*, 283–285.

Mis, F. W., & Moore, J. W. Effect of preacquisition UCS exposure on classical conditioning of the rabbit's nictitating membrane response. *Learning and Motivation*, 1973, *4*, 108–114.

Mitchell, D. Reply to Revusky. *Animal Learning & Behavior*, 1977, *5*, 321.

Mitchell, D., Scott, D. W., & Mitchell, L. K. Attenuated and enhanced neophobia in the taste-aversion "delay of reinforcement" effect. *Animal Learning & Behavior*, 1977, *5*, 99–102.

Nachman, M. Learned aversion to the taste of lithium chloride and generalization to other salts. *Journal of Comparative and Physiological Psychology*, 1963, *56*, 343–349.

Nachman, M. Learned taste and temperature aversions due to lithium chloride sickness after temporal delays. *Journal of Comparative and Physiological Psychology*, 1970, *73*, 22–30.

Nachman, M., & Ashe, J. H. Learned taste aversions in rats as a function of dosage, concentration, and route of administration of LiCl. *Physiology & Behavior*, 1973, *10*, 73–78.

Nowlis, G. H. Conditioned stimulus intensity and acquired alimentary aversions in the rat. *Journal of Comparative and Physiological Psychology*, 1974, *86*, 1173–1184.

Passey, G. E. The influence of intensity of unconditioned stimulus upon acquisition of a conditioned response. *Journal of Experimental Psychology*, 1948, *38*, 420–428.

Pavlov, I. P. *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. London: Oxford University Press, 1946. (Originally published, 1927.)

Petrinovich, L., & Bolles, R. Delayed alternation: Evidence for symbolic processes in the rat. *Journal of Comparative and Physiological Psychology*, 1957, *50*, 363–365.

Petrinovich, L., Bradford, D., & McGaugh, J. L. Drug facilitation of memory in rats. *Psychonomic Science*, 1965, *2*, 191–192.

Pschirrer, M. Goal events as discriminative stimuli over extended intertrial intervals. *Journal of Experimental Psychology*, 1972, *96*, 425–432.

Ray, O. S., & Bivens, L. W. Reinforcement magnitude as a determinant of performance decrement after electroconvulsive shock. *Science*, 1968, *160*, 330–332.

Renner, K. E. Delay of reinforcement: A historical review. *Psychological Bulletin*, 1964, *61*, 341–361.

Rescorla, R. A. Pavlovian conditioned inhibition. *Psychological Bulletin*, 1969, *72*, 77–94.

Rescorla, R. A. Variation in the effectiveness of reinforcement and nonreinforcement following prior inhibitory conditioning. *Learning and Motivation*, 1971, *2*, 113–123.

Rescorla, R. A. Pavlovian second-order conditioning: Some implications for instrumental behavior. In H. Davis & H. M. B. Hurwitz (Eds.), *Operant-Pavlovian interactions*. Hillsdale, N.J.: Erlbaum, 1977.

Revusky, S. H. Aversion to sucrose produced by contingent x-irradiation: Temporal and dosage parameters. *Journal of Comparative and Physiological Psychology*, 1968, *65*, 17–22.

Revusky, S. H. The role of interference in association over a delay. In W. K. Honig & P. H. R. James (Eds.), *Animal memory*. New York: Academic Press, 1971.

Revusky, S. H. The concurrent interference approach to delay learning. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977. (a)

Revusky, S. H. Correction of a paper by Mitchell, Scott, and Mitchell. *Animal Learning & Behavior*, 1977, *5*, 320. (b)

Revusky, S. H. Interference with progress by the scientific establishment: Examples from flavor aversion learning. In N. W. Milgram, L. Krames, & T. M. Alloway (Eds.), *Food aversion learning*. New York: Plenum Press, 1977. (c)

Revusky, S. H. Learning as a general process with an emphasis on data from feeding experiments. In N. W. Milgram, L. Krames, & T. M. Alloway (Eds.), *Food aversion learning*. New York: Plenum Press, 1977. (d)

Revusky, S. H., & Bedarf, E. W. Association of illness with prior ingestion of novel foods. *Science*, 1967, *155*, 219–220.

Revusky, S. H., & Garcia, J. Learned associations over long delays. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 4). New York: Academic Press, 1970.

Revusky, S. H., Parker, L. A., & Coombes, S. Flavor aversion learning: Extinction of the aversion to an interfering flavor after conditioning does not affect the aversion to the reference flavor. *Behavioral Biology*, 1977, *19*, 503–508.

Riccio, D. C., & Haroutunian, V. Failure to learn in a taste aversion paradigm: Associative or performance deficit? *Bulletin of the Psychonomic Society*, 1977, *10*, 219–222.

Riccio, D. C., Rohrbaugh, M., & Hodges, L. A. Developmental aspects of passive and active avoidance learning in rats. *Developmental Psychobiology*, 1968, *1*, 108–111.

Riley, A. L., & Baril, L. L. Conditioned taste aversions: A bibliography. *Animal Learning & Behavior*, 1976, *4*, 1S–13S.

Riley, A. L., & Clarke, C. M. Conditioned taste aversions: A bibliography. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977.

Rizley, R. C., & Rescorla, R. A. Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology*, 1972, *81*, 1–11.

Rozin, P. Central or peripheral mediation of learning with long CS-US intervals in the feeding system. *Journal of Comparative and Physiological Psychology*, 1969, *67*, 421–429.

Rozin, P. The evolution of intelligence and access to the cognitive unconscious. In J. M. Sprague & A. N. Epstein (Eds.), *Progress in psychobiology and physiological psychology* (Vol. 6). New York: Academic Press, 1976.

Rozin, P. The significance of learning mechanisms in food selection: Some biology, psychology, and sociology of science. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977.

Rozin, P., & Kalat, J. W. Specific hungers and poison avoidance as adaptive specializations of learning. *Psychological Review*, 1971, *78*, 459–486.

Rozin, P., & Ree, P. Long extension of effective CS–US interval by anesthesia between CS and US. *Journal of Comparative and Physiological Psychology*, 1972, *80*, 43–48.

Rudy, J. W., & Cheatle, M. D. Odor-aversion learning in neonatal rats. *Science*, 1977, *198*, 845–846.

Rudy, J. W., Iwens, J., & Best, P. J. Pairing novel exteroceptive cues and illness reduces illness-induced taste aversions. *Journal of Experimental Psychology: Animal Behavior Processes*, 1977, *3*, 14–25.

Scarborough, B. B., Whaley, D. L., & Rogers, J. G. Saccharin avoidance behavior instigated by x-irradiation in backward conditioning paradigms. *Psychological Reports*, 1964, *14*, 475–481.

Schwartz, B. On going back to nature: A review of Seligman and Hager's *Biological boundaries of learning*. *Journal of the Experimental Analysis of Behavior*, 1974, *21*, 183–198.

Seligman, M. E. P. On the generality of the laws of learning. *Psychological Review*, 1970, *77*, 406–418.

Seligman, M. E. P. Reply to Malone. *Psychological Review*, 1973, *80*, 306.

Seligman, M. E. P., & Hager, J. L. *Biological boundaries of learning*. New York: Appleton-Century-Crofts, 1972.

Shettleworth, S. J. Constraints on learning. In D. S. Lehrman, R. A. Hinde, & E. Shaw (Eds.), *Advances in the study of behavior* (Vol. 4). New York: Academic Press, 1972. (a)

Shettleworth, S. J. Stimulus relevance in the control of drinking and conditioned fear responses in domestic chicks (*Gallus gallus*). *Journal of Comparative and Physiological Psychology*, 1972, *80*, 175–198. (b)

Siegel, S. Flavor preexposure and "learned safety." *Journal of Comparative and Physiological Psychology*, 1974, *87*, 1073–1082.

Siegel, S., & Domjan, M. Backward conditioning as an inhibitory procedure. *Learning and Motivation*, 1971, *2*, 1–11.

Skinner, B. F. A case history in scientific method. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 2). New York: McGraw-Hill, 1959.

Smith, D. F., & Balagura, S. Role of oropharyngeal factors in LiCl aversion. *Journal of Comparative and Physiological Psychology,* 1969, *69*, 308–310.

Smith, J. C., & Roll, D. L. Trace conditioning with x-rays as an aversive stimulus. *Psychonomic Science,* 1967, *9,* 11–12.

Staddon, J. E. R., & Simmelhag, V. L. The "superstition" experiment: A reexamination of its implications for the principles of adaptive behavior. *Psychological Review,* 1971, *78,* 3–43.

Sutherland, N. S., & Mackintosh, N. J. *Mechanisms of animal discrimination learning.* New York: Academic Press, 1971.

Taukulis, H. K., & Revusky, S. H. Odor as a conditioned inhibitor: Applicability of the Rescorla-Wagner model to feeding behavior. *Learning and Motivation,* 1975, *6,* 11–27.

Testa, T. J., & Ternes, J. W. Specificity of conditioning mechanisms in the modification of food preferences. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection.* Waco, Tex.: Baylor University Press, 1977.

Thorndike, E. L. *The fundamentals of learning.* New York: Bureau of Publications, Teachers College, Columbia University, 1932.

Thorndike, E. L. *Animal intelligence.* New York: Hafner, 1965. (Originally published, 1911.)

Timberlake, W., & Grant, D. L. Auto-shaping in rats to the presentation of another rat predicting food. *Science,* 1975, *190,* 690–692.

Trent, A. M., & Kalat, J. W. Lack of effect of specific sodium hunger on learned aversions to sodium and sucrose. *Animal Learning & Behavior,* 1977, *5,* 243–246.

Tyler, D. W., Wortz, E. C., & Bitterman, M. E. The effect of random and alternating partial reinforcement on resistance to extinction in the rat. *American Journal of Psychology,* 1953, *66,* 57–65.

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology,* 1968, *76,* 171–180.

Wagner, A. R., & Terry, W. S. Backward conditioning to a CS following an expected vs. a surprising UCS. *Animal Learning & Behavior,* 1975, *3,* 370–374.

Weisinger, R. S., Parker, L. F., & Skorupski, J. D. Conditioned taste aversions and specific need states in the rat. *Journal of Comparative and Physiological Psychology,* 1974, *87,* 655–660.

Weisman, R. G. On the role of the reinforcer in associative learning. In H. Davis & H. M. B. Hurwitz (Eds.), *Operant-Pavlovian interactions.* Hillsdale, N.J.: Erlbaum, 1977.

Wilcoxon, H. C. Long-delay learning of ingestive aversions in quail. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection.* Waco, Tex.: Baylor University Press, 1977.

Wilcoxon, H. C., Dragoin, W. B., & Kral, P. A. Illness-induced aversions in rat and quail: Relative salience of visual and gustatory cues. *Science,* 1971, *171,* 826–828.

Zahorik, D. M. Associative and nonassociative factors in learned food preferences. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection.* Waco, Tex.: Baylor University Press, 1977.

Zahorik, D. M., & Houpt, K. A. The concept of nutritional wisdom: Applicability of laboratory learning models to large herbivores. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection.* Waco, Tex.: Baylor University Press, 1977.

Zahorik, D. M., & Johnston, R. E. Taste aversions to food flavors and vaginal secretion in golden hamsters. *Journal of Comparative and Physiological Psychology,* 1976, *90,* 57–66.

Received November 8, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 297-306

Determinacy of Common Factors: A Nontechnical Review

Roderick P. McDonald
Ontario Institute for Studies in Education
Toronto, Canada

Stanley A. Mulaik
Georgia Institute of Technology

Most previous accounts of the factor-score indeterminacy problem have failed to give empirical meaning to alternative factor variables. An empirical interpretation of alternative factor variables is developed in terms of alternative tests of infinite length that include a given core set of variables as a subtest. We show that if the core set can be given the same factor loadings on a factor when analyzed alone or in the context of an infinite domain of variables, then there is just one factor variable in the domain that is a possible factor variable of the core set. If consistent factor loadings cannot be found, there is no factor variable in the domain that is a possible factor of the core set. In the latter case, alternative subdomains of variables may contain alternative possible factors of the core set.

Psychologists who use factor analysis undoubtedly know that the common factor model is considered to have special problems that arise from the indeterminacy of its factor scores. Nevertheless, the actual nature of these special problems is almost certainly not widely understood.

Recent publications (Green, 1976; McDonald, 1974, 1977; Mulaik, 1972, 1976; Mulaik & McDonald, 1978; Schönemann, 1971; Schönemann & Steiger, 1976; Schönemann & Wang, 1972) have dealt with the problem of factor-score indeterminacy in a highly technical manner. Consequently, the psychologist who uses factor analytic techniques occasionally in his research may be confused about the conflicting views that have been expressed on this problem and its meaning for his own research, and he may be unaware of a developing consensus about the problem and its implications.

The objects of this article are first to give a relatively nonmathematical account of factor-score indeterminacy, second to review the major recent discussions of it, and third to offer a resolution of the conflicting views that have arisen, while making clear the implications of factor-score indeterminacy for the definition of common factors and for the practice of factor analytic research.

Factor-score indeterminacy refers to the fact that the common and unique factor scores in the common factor model are not uniquely determined by the observed variables whose correlations they explain, since in general the multiple correlation between a common or unique factor and the observed variables is less than unity.

This fact has sometimes been taken to mean that one cannot obtain exact scores on a common factor and that one must therefore settle for regression estimates of them. Such a view is not strictly correct, as it is quite possible to construct numbers that behave precisely like scores on a given common factor. The difficulty is instead that infinitely many sets of such numbers can be constructed, each set in correspondence with a given set of observations. In other words, the factor scores do not have a unique mathematical definition. Given a factor analysis of the observed variables, with factor pattern, factor structure, and factor correlations all determined to one's

Requests for reprints should be sent to Roderick P. McDonald, Ontario Institute for Studies in Education, 252 Bloor Street, West, Toronto, Ontario, Canada M5S 1V6.

satisfaction (i.e., with rotational indeterminacy eliminated), there are still infinitely many random variables that can satisfy the conditions for being a possible factor variable that corresponds to each column of factor loadings, and there need not be a high correlation between two alternative possible factor variables.

The observation that factor scores lack a unique mathematical definition has led some writers to conclude that there is something wrong with the common factor model and that we should abandon it in favor of some approximately equivalent model that is not also believed to have something wrong with it. The usual alternative is some version of component theory (e.g., Kaiser, 1970; Schönemann & Steiger, 1976).

On the other hand, many psychologists who use factor analysis would not see any importance in the questions that have been raised about the indeterminacy of factor scores. Essentially, they know very well how seldom they require the assessment of factor scores, and they see no logical connection between questions about factor scores and the widely accepted aim of factor analysis, namely, the interpretation of a common factor in terms of the common attribute of the tests that have high factor loadings on it. (Indeed, so strong is the effect of this practice that some users forget basic theory and come to think of a factor as a profile of factor loadings of tests, not as a score that characterizes a subject.)

It turns out, however, that in exploratory factor analysis as it is usually employed, the range of possible mathematical constructions of a possible factor variable corresponds to a range of possible ambiguity in the interpretation of a common factor. Moreover, this ambiguity cannot be avoided by approximating the model with components. In its broader implications, factor-score indeterminacy does not merely concern the problem of obtaining a score, or the problem of obtaining too many scores, that might be the score of an individual on a common factor. Rather, it concerns the inability of a finite set of observed variables in an exploratory factor analysis to determine unambiguously what attribute of the individuals the factor variable represents. This is important, as one will see, because factor analysis has commonly been treated as a theory-generating device; that is, it has been treated as a device for the post facto discovery of the psychological concepts that explain the correlations of the variables one has chosen to measure.

In the following discussion, we simplify the issues by concentrating on the special case of common factor analysis in which there is just one general factor (the Spearman case). In this special case, the mathematics can be kept simple, and there is no possible confusion between the indeterminacy of factor scores and the indeterminacy of factor loadings due to rotation. For the most part, the corresponding results for multiple factors are straightforward analogues.

The Basic Results

We begin by considering a vector of n random variables, $\mathbf{y}' = (Y_1, \ldots, Y_n)$, each of which has a mean of zero and a standard deviation of unity. In most applications one can think of $\mathbf{y}$ as corresponding to the scores on n tests obtained by a person randomly drawn from an infinite population $\mathcal{P}$.

A very simple and sufficient definition of the general factor model is to postulate that there exists a random variable X (an additional variable defined on the set of persons in $\mathcal{P}$) such that the n (linear) regression estimators of $Y_1, \ldots, Y_n$ from X have mutually uncorrelated residuals. This defines the model completely. We express this statement symbolically by writing

$$\hat{\mathbf{y}} = \mathbf{f}X, \tag{1}$$

where $\hat{\mathbf{y}}' = (\hat{Y}_1, \ldots, \hat{Y}_n)$, $\mathbf{f}' = (f_1, \ldots, f_n)$, and $\hat{Y}_j$ is the regression estimator of Y_j from X with slope f_j. The vector $\mathbf{e}$ of residuals is then defined by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}, \tag{2}$$

and by definition the residuals are uncorrelated with X. It then follows that the $(n \times n)$ correlation matrix $\mathbf{y}$ takes the form

$$\mathbf{R} = \mathbf{f}\mathbf{f}' + \mathbf{U}, \tag{3}$$

where by the assumption of uncorrelated residuals,

$$\mathbf{U} = \text{diag}\,(1 - f_1^2, \ldots, 1 - f_n^2). \tag{4}$$

In conventional terminology, the regression weights f_j are *factor loadings*. These are also the correlations of the tests with the general factor. The residual variances $u_{jj} = 1 - f_j^2$ are the *unique variances*.

In applications of the model, one attempts to fit the hypothesis in Equation 3 to a sample correlation matrix. If the fit is acceptable, one interprets the general factor X as whatever attribute of the persons in $\mathcal{P}$ seems to be common to the tests.

It is reasonable to suppose that all unique variances are greater than zero, that is, no test is perfectly correlated with the factor, if only because of errors of measurement. If these assumptions are true and the number of tests is finite, the multiple correlation of the common factor X with the observed variables is strictly less than unity. The regression estimator of X from $\mathbf{y}$ is given by

$$\hat{X} = \mathbf{f}'\mathbf{R}^{-1}\mathbf{y}, \tag{5}$$

and the corresponding squared multiple correlation ρ^2 of X and $\hat{X}$ is given by

$$\rho^2 = \mathbf{f}'\mathbf{R}^{-1}\mathbf{f}. \tag{6}$$

Because of Equation 4, Equation 5 can be expressed as

$$\hat{X} = \frac{1}{g}\sum_{j=1}^{n} \frac{f_j^2}{1 - f_j^2} Y_j, \tag{7}$$

and Equation 6 can be expressed as

$$\rho^2 = g/g + 1, \tag{8}$$

where

$$g = \sum_{j=1}^{n} [f_j^2/(1 - f_j^2)] \tag{9}$$

(see, e.g., McDonald & Burr, 1967). From the form of Equation 8, if g is finite, ρ^2 is strictly less than unity (since g is strictly less than $g + 1$). If one can conceivably find infinitely many variables with nonzero loadings on X, then in the limit as n approaches infinity, g approaches infinity and

$$\lim_{n\to\infty} \rho^2 = 1 \tag{10}$$

(Piaggio, 1931, 1933). One can think of Equation 8 with Equation 9 as a generalization of the Spearman–Brown formula for the reliability of a test as a function of its length (cf. McDonald & Burr, 1967, p. 392). In effect, one sees that the general factor cannot be determined with perfect reliability by a test battery of finite length. This simple fact is not, however, the problem of factor-score indeterminacy. If it were, one would deal with it in practice by simply admitting that one expects all social science measurements to be imperfect, and one would be satisfied with values of ρ^2 in the general range of acceptable reliability coefficients. These values are ordinarily obtainable in factor analytic studies.

Instead the problem of factor-score indeterminacy as it has been presented is as follows. A necessary and sufficient condition for a random variable X^* (in standard score form) defined over the population $\mathcal{P}$ to have the properties of X in Equations 1–4 is that the correlations of X^* with $Y_1, \ldots, Y_n$ be given by

$$\rho(X^*, Y_j) = f_j, \quad \text{where } j = 1, \ldots, n. \tag{11}$$

If a standardized random variable X^* has the required correlation with $\mathbf{y}$, one says that it is a possible factor variable of $\mathbf{y}$. There are infinitely many such factor variables definable over $\mathcal{P}$ (Guttman, 1955; McDonald, 1974; Mulaik, 1976). (Note also that $\hat{X}$ is not itself a possible factor variable of $\mathbf{y}$.)

If X_1 and X_2 are two standardized random variables that have the required correlations with $\mathbf{y}$, one can obtain bounds on the correlation ρ_{12} between them by showing that the partial correlation of X_1 and X_2, with $Y_1, \ldots, Y_n$ partialed out, is given by

$$\rho_{12\cdot\mathbf{y}} = \rho_{12} - \rho^2/1 - \rho^2. \tag{12}$$

Since this partial correlation must lie between -1 and $+1$, it follows that

$$2\rho^2 - 1 \leq \rho_{12} \leq 1, \tag{13}$$

where ρ^2 is given by Equation 8 as before. This result was first given by Guttman (1955) in the more general case of multiple factors, using a different line of argument.

Another necessary and sufficient condition for X^* to have the properties of the general

factor X in Equations 1–4 is that

$$X^* = \hat{X} + D^*, \qquad (14)$$

where $\hat{X}$ is the regression estimator given by Equation 7 and D^* is any random variable defined over the population that has the properties of a residual of X about $\hat{X}$, that is, any random variable that has variance $1 - \rho^2$ and correlations

$$\rho(D^*, Y_j) = 0, \quad \text{where } j = 1, \ldots, n. \qquad (15)$$

This result was first given by Spearman (1922). It was also obtained by Heywood (1931) rather more rigorously and, in the general case of multiple factors, by Kestelman (1952) and Guttman (1955). (The latter was the first to establish that the condition is necessary as well as sufficient.)

That there is a problem of factor-score indeterminacy, beyond the mere fact of imprecision of the estimates given by $\hat{X}$, is first noticed when one observes that since the range of possible values of ρ^2 is from zero to unity, the range of possible minimum values of ρ_{12} is from minus unity to unity; and if ρ is less than $1/(2)^{\frac{1}{2}}$, then ρ_{12} is possibly zero. That is, if the correlation between the estimator and a possible factor variable is less than .707, the correlation between two such possible factor variables can be as low as zero. It is not immediately obvious what empirical interpretation to place on this fact, however.

Interpretations of Factor-Score Indeterminacy

Guttman (1955), having introduced the bounds (Inequality 13) and having noted in particular that one requires $\rho^2 > .71$ if the lower bound on ρ_{12} is to be positive, set out its implications for the scientific meaning of factor analysis in the following remarks:

It appears from the relation of $[2\rho^2 - 1]$ to $[\rho^2]$ (for each intended factor) that the predictability of factors from [y] is not merely a practical problem. If $[2\rho^2 - 1]$ is low, it raises the question of what it is that is being estimated in the first place; instead of only *one* 'primary' trait there are many widely different variables associated with a given profile of loadings. . . . If more direct observations on factor scores cannot be made than statistical analysis of R and [y], the Spearman–Thurstone approach may have to be discarded for lack of determinacy of its factor scores. (p. 79)

Similarly one finds:

The widespread practice of trying to name or attach meaning to factors merely by studying factor *loadings* is clearly suspect if the same loadings can be derived equally well from radically different sets of factor scores. (Guttman, 1957, p. 149).

Schönemann and Wang (1972) pointed out that in estimating the factor model on the Lawley–Rao basis (which is usually employed in fitting the unrestricted model by maximum likelihood), instead of retaining factors corresponding to eigenvalues greater than unity, one would have to retain only factors with eigenvalues greater than two

if one were to insist on factors which are better determined than a set of standardized random numbers are (in the sense that the scores on factor $[X_1]$ can be predicted better from scores on its equivalent twin $[X_2]$ than from a set of random numbers. (Schönemann & Wang, 1972, p. 69)

They then proceeded to show that in empirical studies, some factors that are needed to explain the correlations do not have acceptable—apparently meaning positive—lower bounds to the correlation between possible alternative factors, and they claimed that this appears to create a dilemma for the user of the model.

We notice a difficulty in coming to grips with these interpretations of factor-score indeterminacy. As stated above, given at least the existence of errors of measurement, no test of finite length will be perfectly correlated with the factor, hence no test of finite length will have the properties of a common factor variable. It is not enough to be told that attaching meanings to factors on the basis of factor loadings is "clearly" rendered suspect by the existence of alternative possible factor variables. One needs also to be told how to find at least two distinct possible factor variables and on what grounds one might place distinct interpretations on them. Instead, Guttman (1955) showed only that one can construct alternative factor variables in infinitely many ways by adding to a regression estimate $\hat{X}$ a value of a random variable D^*, as in Equations 14 and 15, defined "by throwing dice or turning a roulette wheel" (p. 70). More precisely, to construct a possible factor score for any person randomly drawn from the defined population $\mathcal{P}$, one makes one

throw of the dice, or turn of the roulette wheel (or generates one number by means of a random number generator), and arbitrarily associates the additional random number D^* so obtained with the scores of the randomly drawn person. Schönemann and Wang (1972) suggested similar procedures for constructing possible factor scores and remarked that since infinitely many different sets of such numbers can be computed, they need not be estimated.

However, although such computational procedures do indeed yield mathematically admissible alternative factor variables, the solutions so obtained can hardly be regarded as measurements of empirical properties of persons in the population $\mathcal{P}$. Measurement constitutes the assignment of numbers to objects in such a way as to represent empirical relationships by numerical relationships. It is hard to see how assigning artificially generated random numbers to persons in the population can represent empirical relationships (of order, say) among these persons. Thus, these alternative factor variables lack empirical significance and as such cannot be regarded as lending themselves to distinct interpretations of the factor. Hence the argument that factor indeterminacy renders the interpretation of factor loadings suspect is incomplete. Therefore, it is understandable if the typical user of factor analysis chooses to disregard this argument.

McDonald (1974) sought to offer an escape from the mathematical dilemma pointed out by Schönemann and Wang (1972). Based on a general mathematical theorem about the construction of and relationship between all possible factor variables, McDonald's discussion consisted essentially of two main arguments.

The first of these was a demonstration that the correlation between two distinct simulations of a factor variable, using independent residual variables D^* in Equation 14 (generated by using the artificial devices suggested by Guttman, 1955, and by Schönemann and Wang, 1972), is ρ^2. Two implications of this are (a) that ρ^2 is the correlation between any possible factor of y, defined (without simulation) over $\mathcal{P}$, and any simulation of it created by the devices recommended and (b) that ρ^2 is the correlation between two independent simulations of a possible factor of y made by independent investigators.

The second argument amounted to a declaration that for consistency one must adopt the convention that the common factor be thought of as uniquely defined, though unobservable; hence the (minimum) correlation between alternative possible factor variables cannot serve as a measure of the indeterminacy of the common factor, since the distinct values of these alternatives for a given person in $\mathcal{P}$ cannot both be the unique but unknown value of his factor.

The first of McDonald's arguments is correct, but is not, as it turns out, particularly useful. The second, it can now be shown, is ambiguous and incomplete and seemingly has proved unconvincing.

Mulaik (1976) rejected McDonald's second argument, claiming that Guttman's lower bound represents a measure of the extent of possible disagreement between two investigators about the nature of a factor. Mulaik's argument assumes that two investigators might actually find two distinct empirical measures, either of which has the properties of the common factor variable. Further, the empirical measures might be imperfectly correlated, and their test contents might give conflicting interpretations of the factor (presumably whether or not these were consistent with the contents of the remaining n tests). Mulaik further argued that in typical applications factor analysts will have at most only a few alternative empirical interpretations of a factor and that the expected correlation between pairs of alternative factor variables (across a universe of such pairs, randomly selected) will be greater than or equal to ρ^2 in Equation 6.

Green (1976) offered a generally conciliatory conclusion on what he described as the "factor indeterminacy controversy," suggesting that all possible factor variables are equally true and that both ρ^2 and $2\rho^2 - 1$ serve as measures of factor-score indeterminacy.

With the exception of remarks in Mulaik (1976), what all of these discussions have in common is a failure to relate the problem described to any discernible realities of factor analytic practice.

It turns out that although it is not true that

different investigators might construct different empirical measures of finite length that satisfy Inequality 13 and hence are possible factor variables, there are conditions under which they might construct the beginnings of sequences of variables (each of which contains at least some error of measurement) that would in the limit yield tests of infinite length that satisfy Inequality 13. This fact and its consequences form the topic of the next section.

Determinacy of Common Factors in a Behavior Domain

One way to give empirical meaning to alternative possible common factor variables has been shown by Mulaik and McDonald (1978). They considered the possible factors of two sets of variables obtained by augmenting a core set of n variables (*marker* variables for a factor, one might say) with different sets of additional variables, subject only to the condition that each augmented set continue to satisfy a common factor model consistent with the model for the core set. They discussed what one might call alternative tests of infinite length, developed from the given core set and each uniquely determining a factor variable. They gave conditions under which these tests of infinite length will determine the same factor variable. If these conditions are not satisfied, the tests may determine distinct factor variables whose correlations in the limit are subject only to the Guttman bounds (Inequality 13). See also McDonald (1977).

Here we give a simpler treatment of the problem, directly based on behavior domain theory. Following Guttman (1957), we suppose that the n variables $Y_1, \ldots, Y_n$ of the section entitled The Basic Results are drawn from a set of $n + m$ empirical variables $Y_1, \ldots, Y_n$ and $Y_{n+1}, \ldots, Y_{n+m}$, a *universe of content* or *behavior domain* that is defined in advance of any statistical analysis.

An ambiguity concerning the possible alternative factor variables of $Y_1, \ldots, Y_n$ is removed by supposing that $f_1, \ldots, f_n$ are fixed, known values. In practice, this means that one must have at least three variables in the core, whose correlations then determine the factor loadings uniquely. (In the multiple-factor counterpart of these arguments, we would require that the factor loadings be fixed against rotation and known.)

Write

$$\mathbf{y}'_1 = (Y_1, \ldots, Y_n), \quad \mathbf{y}'_2 = (Y_{n+1}, \ldots, Y_{n+m}),$$

and

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} \tag{16}$$

for the joint correlation matrix of $\mathbf{y}_1$ and $\mathbf{y}_2$, partitioned conformably.

One cannot seek to establish any relationship between the possible factor variables of $Y_1, \ldots, Y_n$ and those of the behavior domain, except under the condition that one is able to obtain a joint factor analysis of the $n + m$ variables in which the factor loadings of the first n are the same as when the first n is factored without the further m. If this condition is satisfied, we say that the core and the domain have consistent loadings. If the core and the domain do not have consistent loadings, there is no more reason to seek a relationship between their possible factors than to seek a relationship between two of the multiple factors in an orthogonal multiple factor analysis. Although the effect of adding more variables may be to add further common factors, so that, say, there are r common factors altogether, the requirement of consistent loadings is that one must be able to fit the model and choose a rotation in such a way that the factor loadings of the n core variables are unaltered in the context of the additional m variables. It can hence be seen that the core and the domain are consistent in this sense if and only if the factor solution for the domain is of the form

$$\begin{aligned} &\begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{f}_1 & \mathbf{G}_1 \\ \mathbf{f}_2 & \mathbf{G}_2 \end{bmatrix} \begin{bmatrix} \mathbf{f}'_1 & \mathbf{f}'_2 \\ \mathbf{G}'_1 & \mathbf{G}'_2 \end{bmatrix} + \begin{bmatrix} \mathbf{U}_1 & \\ & \mathbf{U}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{f}_1\mathbf{f}'_1 + \mathbf{D} + \mathbf{U}_1 & \mathbf{f}_1\mathbf{f}'_2 + \mathbf{G}_1\mathbf{G}'_2 \\ \mathbf{f}_2\mathbf{f}'_1 + \mathbf{G}_2\mathbf{G}'_1 & \mathbf{f}_2\mathbf{f}'_2 + \mathbf{G}_2\mathbf{G}'_2 + \mathbf{U}_2 \end{bmatrix}, \end{aligned} \tag{17}$$

where

$$\mathbf{G}_1\mathbf{G}'_1 = \mathbf{D}, \tag{18}$$

a diagonal matrix with nonnegative diagonal terms, and

$$\mathbf{U}_1 = \mathbf{U} - \mathbf{D}, \qquad (19)$$

which also contains nonnegative diagonal terms and where $\mathbf{U}$ is the original unique variance matrix given by Equation 4. What this means is that the effect of adding variables may be to redefine some of the unique variance of one or more components of $\mathbf{y}_1$ as common variance, that is, variance in common with that of components of $\mathbf{y}_2$. The loadings in $\mathbf{G}_1$, with the property defined in Equation 18, take care of this possibility, but do not account for any correlation between components of $\mathbf{y}_1$. The loadings in $\mathbf{G}_2$ account for some of the correlation between components of $\mathbf{y}_2$ and, together with $\mathbf{G}_1$, for any correlation between components of $\mathbf{y}_1$ and components of $\mathbf{y}_2$ that is not accounted for by the original general factor.

One now recognizes a simple result of fundamental importance: A possible common factor variable of the domain is a possible general factor of the core factored separately, if and only if the domain and the core have consistent loadings. To see this, note that if one has the structure of Equation 17 with Equation 18 and 19, one can write the corresponding factor model in the form

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{G}_1 \\ \mathbf{f}_2 & \mathbf{G}_2 \end{bmatrix} \begin{bmatrix} X \\ \mathbf{z} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}, \qquad (20)$$

where $\mathbf{z}$ is a vector of $r - 1$ additional common factor variables and the residuals $\mathbf{e}_1$ and $\mathbf{e}_2$ have covariances given by $\mathbf{U}_1$ and $\mathbf{U}_2$. Then if $\mathbf{y}_2$ is deleted, Equation 20 become

$$\mathbf{y}_1 = \mathbf{f}_1 X + \mathbf{G}_1 \mathbf{z} + \mathbf{e}_1 = \mathbf{f}_1 X + \mathbf{e}, \qquad (21)$$

where $\mathbf{e}$ is given by Equation 2, since $\mathbf{G}_1\mathbf{z} + \mathbf{e}_1$ has the covariance matrix $\mathbf{U} = \mathbf{U}_1 + \mathbf{D}$ and behaves just like the original residual. Conversely, if there is a common factor of the joint set that is a possible general factor of the core set (possibly with additional common factors), then it satisfies Equation 11, and so it must be possible to express the joint common factor pattern in the required form. This completes the proof of the stated result

Further, by an obvious extension of the well-known result (Equation 10) given by Piaggio (1931, 1933), it follows that if m, the number of additional variables in the domain, is infinitely large, then the possible first factor is uniquely determined by the $n + m$ variables that constitute the domain, since its multiple correlation with them is unity. (This assumes that a nonzero proportion of the components of $\mathbf{f}_2$ is strictly greater than zero.)

Thus one finds that if the domain factor pattern is not consistent with that of the core, that is, if it cannot take the form of Equation 17 with Equations 18 and 19 and the conditions on them, then there is no possible common factor variable of the domain that is also a possible general factor variable of the core set. If the joint factor pattern is consistent with the core set, so that possible factor variables of the domain and of the core set can be defined on the same basis, and if the additional variables of the domain have nonzero loadings on the factor that is marked by the core, there is one and only one factor variable in the domain that is a possible factor of the core set. In this case, the factor loadings of the core set mark a unique factor variable in the domain, which is determined as precisely as one pleases by the addition of further variables with nonzero loadings. This confirms the conclusion given by McDonald (1974) on purely formal grounds. In other words, there are not infinitely many possible alternative factor variables in a behavior domain that correspond to the profile of loadings of a given core set of variables from the domain. Either there is just one factor variable in the domain consistent with the core or there is no such factor variable.

If there is a factor variable of the domain that is consistent with the core, it can be shown that the squared correlation between it and the estimator given by Equation 5 is ρ^2 (given by Equation 6), as one would expect. McDonald (1974) has shown that the index ρ^2 as a measure of the determinacy of factor scores does not yield any of the dilemmas associated with the lower bound, $2\rho^2 - 1$.

However, before one concludes that the problem of factor-score indeterminacy has just been eliminated, one must examine the consistency conditions more closely. Essentially, the requirement is that if, in traditional terminology, one "extracts" the first factor,

then the first residual matrix,

$$R^{(1)} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} - \begin{bmatrix} f_1f'_1 & f_1f'_2 \\ f_2f'_1 & f_2f'_2 \end{bmatrix}, \quad (22)$$

must be at least nonnegative definite, that is, a possible covariance matrix of real variables, so that it can be factored in its turn in terms of real factor loadings and nonnegative unique variances. It is important to recognize that there is no reason whatsoever to suppose that this condition will always, or even usually, be fulfilled in practice. Actually, it is a generalized counterpart of the condition that there be no Heywood variables (i.e., variables with negative unique variance). One knows that the condition has been violated if one attempts to add variables to a core, prescribing a consistent analysis, and one obtains a Heywood case. But generally one will not know whether the domain contains variables that, if one were to add them to the core, would violate the required condition.

Mulaik and McDonald (1978) have shown that if a behavior domain does not satisfy conditions similar to Equations 17–19, then such a domain can possibly be divided into overlapping subdomains, all containing the core set as a subset and possessing disjoint sets of factor variables with possible correlations that are only bounded as in Inequality 13. That is, if there is no factor variable in the domain that is a possible factor variable of the core set, there may be factor variables of subdomains that are distinct possible factors of the core set, with correlations anywhere between $2\rho^2 - 1$ and 1. As a consequence, if two investigators take a common core of variables as markers for a factor and independently draw variables from such a domain, not by a sampling rule but by merely augmenting the core set subject to the satisfaction of the general factor model, they can define two distinct tests of infinite length, containing distinct variables that lend themselves to distinct interpretations. Only when the two sets of augmenting variables are merged will one discover that the domain does not have a basis consistent with the core and hence has no factor variable that is a possible factor of the core.

This implies also that one certainly cannot use the factor model to define a behavior domain; that is, a behavior domain is not uniquely defined by choosing a core set of variables that fits the common factor model and then saying that the domain consists of all variables that, jointly with the core set, fit the common factor model with the same number of factors.

Thus, we have arrived at an empirical treatment of the significance of factor indeterminacy. It is a variant of results given by Mulaik & McDonald (1978) and McDonald (1977). If a behavior domain can be factored so that one of its common factors has the same loadings on a core set of variables as when the latter are factored separately, it is then meaningful to consider the relationship between the possible factors of the core and the possible factors of the behavior domain. In such a case, the loadings of the core set uniquely mark a factor variable in the domain, and the addition of further variables with nonzero loadings on the factor will determine it as precisely as one pleases, ultimately yielding an infinite sequence of variables that determines the factor exactly. If the behavior domain cannot be factored consistently with the core, there is no factor variable in the domain that is a general factor of the core set, but there may be many factor variables of subdomains that are distinct alternative general factor variables of the core set. Incompatible sets of variables added to the core set may yield distinct tests of infinite length, defining distinct factor variables and susceptible to distinct interpretations on the basis of their profiles of factor loadings on the distinct tests.

These observations eliminate the paradoxes to be found in the previous purely mathematical treatments of the topic, since the condition under which two possible tests of infinite length, obtained by augmenting a core set, can be correlated zero when each is estimated with $\rho = .707$ is the condition that a corresponding factor in the domain taken as a whole does not exist. If a factor in the domain is defined by the loadings of the core tests, it is defined uniquely.

However, these results do not define away the problem of factor indeterminacy. Rather, they point to its significance for the user of the model. Most users of factor analysis

employ the model in an exploratory fashion. In doing so, they do not begin with a set of relatively clear psychological concepts and then select variables for their study that on the basis of theory should contain or be determined by these concepts. Nor, on the other hand, do they begin with a clearly defined behavior domain (from which all investigators would recognize equally how to develop representative measures) and then select from it a reasonable number of variables to represent the domain in a balanced fashion. Instead, they begin with a number of variables in which they are interested, which might or might not correspond to a general preconception of a behavior domain of tests that "resemble" the tests chosen. Their hope is that when these variables are subjected to a factor analysis, the psychological attributes that determine the correlations among the variables will reveal themselves. Thus, exploratory factor analysis is typically employed to generate a theory rather than to test a theory. It is noticeable that in many such applications the user appears quite satisfied with the results. The problem remains, however, of a range of ambiguity with respect to the ways in which the results may be extended by the use of further variables to measure the discovered attribute more precisely.

If the variables have been selected on the basis of certain common attributes defined in advance, there is no uncertainty as to the interpretation of a factor found common to those variables and no uncertainty as to the further variables one would expect to be associated with the same factor. But in the absence of an agreed domain of variables based on prespecified common attributes, two investigators may seize upon distinct sets of common attributes and proceed to build distinct extended batteries of variables on the basis of the same core set, in terms of these distinct attributes. Ultimately, in principle, they can create two distinct test batteries of infinite length, built on the same original test and corresponding to two distinct attributes and to two distinct random variables, either of which might have been the common factor of the original variables. It is factor indeterminacy that supplies a mathematical latitude for such sets of alternative common attributes to be found. This does not mean that they will inevitably be found in practice. It means only that such possibilities exist. It remains for appropriate empirical work, using merged variables, augmented sets, and extension procedures (see McDonald, 1978), to establish how serious these problems are in practice.

What this analysis reveals, essentially, is that it may be naive to use exploratory factor analysis to generate theory out of an ill-defined domain of variables. The danger is that the factor structure of the entire domain may not be remotely consistent with the factor structure of its subdomains.

However, it must be pointed out that these problems of the common factor model do not, as is sometimes supposed, constitute grounds for using alternative modes of analysis such as component theory. McDonald (1975) has discussed some aspects of this choice. Here one recognizes that if one worries about the relationship of factor-score estimates to the variables being estimated, it is because one is able, under certain conditions, to define the variables being estimated on the basis of a behavior domain that fits the common factor model consistently with the given variables. To show that component scores do not have an indeterminacy problem, as studied here, one must show that the component score of a core set of variables uniquely determines a corresponding component score in an infinite domain from which the core set has been drawn. But one knows immediately that this cannot be true. Even if one could find a way to define consistent bases for the core and the domain, and this seems unlikely, the domain component scores would contain error parts and specific parts that would be uncorrelated with variables in the core. Thus, it seems that mathematical arguments (e.g., Schönemann and Steiger, 1976) that favorably contrast component theories, whether principal components, components of images, or other components, with the common factor model cannot be accepted until they are reformulated in terms of behavior domains and in terms of the relation of core components to domain components. It seems unlikely that core components can be shown to yield better representations of components of interest in a

behavior domain than we have in common factor estimates. (See McDonald, 1977, for a more extended account of this question.)

The implications of factor indeterminacy also render suspect the frequently used technique of including sets of marker variables in distinct sets of additional variables, in the hope of determining whether the same factors are to be found in these distinct sets. It is supposed that if the marker variables have the same loadings on a common factor in each of these sets, this is evidence that the same factor variable has turned up in each of these sets. But this need not be so. The marker variables may be associated with different factor variables when embedded within the contexts of different sets of additional variables. The only way to be sure that the same factors are common across the sets is to analyze the marker variables jointly with all the other variables to see if one can obtain consistent loadings. But if this is done, then there is no need for marker variables in the first place.

Because of the close parallels between common factor theory and classical true score theory, factor indeterminacy has important implications for test theory and for test construction in practice. For example, in the construction of homogeneous tests, one might begin with a core set of items that are thought to define a dimension of interest and then add to that core additional items that are factorially consistent with the original core to increase the total-score reliability of the test. If the developer of the tests does this without taking account of the empirical content of the additional items, which in some instances may differ from that of the core variables, he may end up with a test whose total score measures an attribute that is somewhat different from that measured by the original core items. This may happen if the core set of items is selected from one domain with which the core is factorially consistent while the additional items are selected from another domain with which the core items are also factorially consistent. Jointly the union of the two domains of items may not be factorially consistent with the core items, although the domains are factorially consistent separately.

References

Green, B. F. On the factor score controversy. *Psychometrika*, 1976, *41*, 263–266.

Guttman, L. The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology*, 1955, *8*, 65–81.

Guttman, L. Simple proofs of relations between the communality problem and multiple correlation. *Psychometrika*, 1957, *22*, 147–157.

Heywood, H. B. On finite sequences of real numbers. *Proceedings of the Royal Society, Series A*, 1931, *134*, 486–501.

Kaiser, H. F. A second generation little jiffy. *Psychometrika*, 1970, *35*, 401–415.

Kestelman, H. The fundamental equation of factor analysis. *British Journal of Psychology, Statistical Section*, 1952, *5*, 1–6.

McDonald, R. P. The measurement of factor indeterminacy. *Psychometrika*, 1974, *39*, 203–222.

McDonald, R. P. Descriptive axioms for common factor theory, image theory and component theory. *Psychometrika*, 1975, *40*, 137–152.

McDonald, R. P. The indeterminacy of components and the definition of common factors. *British Journal of Mathematical & Statistical Psychology*, 1977, *30*, 165–176.

McDonald, R. P. Some checking procedures for extension analysis. *Multivariate Behavioral Research*, 1978, *13*, 319–325.

McDonald, R. P., & Burr, E. J. A comparison of four methods of constructing factor scores. *Psychometrika*, 1967, *32*, 381–401.

Mulaik, S. A. *The foundations of factor analysis*. New York: McGraw-Hill, 1972.

Mulaik, S. A. Comments on "The measurement of factorial indeterminacy." *Psychometrika*, 1976, *41*, 249–262.

Mulaik, S. A., and McDonald, R. P. The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika*, 1978, *43*, 177–192.

Piaggio, H. T. H. The general factor in Spearman's theory of intelligence. *Nature*, 1931, *127*, 56–57.

Piaggio, H. T. H. Three sets of conditions necessary for the existence of a *g* that is real and unique except in sign. *British Journal of Psychology*, 1933, *24*, 88–105.

Schönemann, P. H. The minimum average correlation between equivalent sets of uncorrelated factors. *Psychometrika*, 1971, *36*, 21–30.

Schönemann, P. H., & Steiger, J. H. Regression component analysis. *British Journal of Mathematical & Statistical Psychology*, 1976, *29*, 175–189.

Schönemann, P. H., & Wang, Ming-Mei Some new results on factor indeterminacy. *Psychometrika*, 1972, *37*, 61–91.

Spearman, C. Correlation between arrays in a table of correlations. *Proceedings of the Royal Society, Series A*, 1922, *101*, 94–100.

Received November 9, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 307–324

In-Group Bias in the Minimal Intergroup Situation: A Cognitive–Motivational Analysis

Marilynn B. Brewer
University of California, Santa Barbara

Experimental research on intergroup discrimination in favor of one's own group is reviewed in terms of the basis of differentiation between in-group and out-group and in terms of the response measure on which in-group bias is assessed. Results of the research reviewed suggest that (a) factors such as intergroup competition, similarity, and status differentials affect in-group bias indirectly by influencing the salience of distinctions between in-group and out-group, (b) the degree of intergroup differentiation on a particular response dimension is a joint function of the relevance of intergroup distinctions and the favorableness of the in-group's position on that dimension, and (c) the enhancement of in-group bias is more related to increased favoritism toward in-group members than to increased hostility toward out-group members. The implications of these results for positive applications of group identification are discussed.

In 1906, sociologist William Sumner articulated a functionalist approach to the nature of intergroup attitudes in his exposition of the concept of ethnocentrism. The differentiation of peoples into distinct ethnic groups originates, according to Sumner, in context of the "conditions of the struggle for existence." At the individual level, the psychological consequences of this differentiation both reflect and sustain the basic state of conflict between the in-group (or "we-group") and out-groups (or "others-groups"):

> The insiders in a we-group are in a relation of peace, order, law, government, and industry, to each other. Their relation to all outsiders, or others-groups, is one of war and plunder. . . . Sentiments are produced to correspond. Loyalty to the group, sacrifice for it, hatred and contempt for outsiders, brotherhood within, warlikeness without—all grow together, common products of the same situation. (Sumner, 1906, p. 12)

From this perspective, then, attitudinal and perceptual biases in favor of members of one's own group over members of other groups are the product of intergroup competition, serving the dual functions of preserving in-group solidarity and justifying exploitation of out-groups. Presumably also, the greater the intensity of competitive interdependence between groups, the more attraction within the in-group and corresponding hostility toward the other group, whereas low levels of interdependence between groups should be associated with relatively little contrast in attitudes toward members of the in-group and out-group (LeVine & Campbell, 1972).

The functionalist concept of intergroup relations is epitomized in the ambitious field experiment undertaken by Sherif and his colleagues in the context of a boys' summer camp (Sherif, Harvey, White, Hood, & Sherif, 1961). In the fully implemented version of the study, conducted in 1954, two groups of 11-year-old boys were formed in isolation from each other for a period of 8 days before being brought into contact under conditions designed to maximize competition and mutual frustration. The resulting intergroup hostility was documented with anecdotal evidence based on observation of overt behavior, supplemented by controlled measures of sociometric preferences, evaluative trait ratings, and estimates of performance by group members during a competitive task. On each of

---

Requests for reprints should be sent to Marilynn B. Brewer, Department of Psychology, University of California, Santa Barbara, California 93106.

these indicators, campers revealed consistent biases favoring members of their own group over members of the competing group. Reductions in bias were not achieved until the nature of the functional relationship between the groups was altered by systematic introduction of "superordinate goals" requiring cooperative interaction.

The Sherif et al. field study is essentially a demonstration rather than a test of the functionalist view of intergroup relations, since its design took for granted that interaction under competitive conditions was prerequisite to the initial development of in-group bias and intergroup hostility. No systematic assessment of attitudes toward in-group and out-group members was made before the intergroup competition phase of the experiment (although changes were documented after competitive pressures had been removed). However, some anecdotal evidence from the 1954 study was provided that indicates that negative reactions to the out-group were present prior to the introduction of structured competition. At the close of the first phase of the experiment, the two groups were first made aware of each other's existence, and at that time the mere knowledge of the presence of the other group was sufficient to generate name-calling and other derogatory commentary from each group directed toward the other (Sherif et al., 1961, p. 95).

The significance of these initial-contact effects has been realized only recently as the phenomena associated with intergroup perception have been reexamined in light of more general cognitive processes by which human beings structure, simplify, and give meaning to their physical and social environment (Hamilton, 1976; Hensley & Duval, 1976; Tajfel, 1969, 1970). From this perspective, any categorization rule that provides a basis for classifying an individual as belonging to one social grouping as distinct from another can be sufficient to produce differentiation of attitudes toward the two groups, in the absence of any initial competitive interdependence. The present review focuses on research directed toward identifying the minimal conditions necessary to generate in-group–out-group discrimination.

Defining the Minimal Intergroup Situation

A number of laboratory studies have attempted to demonstrate the presence of in-group favoritism under conditions in which the independence of outcomes for in-group and out-group is explicitly controlled. Among the earliest of such demonstrations was a study reported by Ferguson and Kelley (1964) in which two groups of three to six members each worked independently on three tasks. Following the interaction group members were asked to rate the quality of the products of both groups separately on a 9-point scale. Mean ratings obtained were significantly biased in the direction of more positive evaluation of subjects' own group's product than of the other group's product, irrespective of any objective differences in output between the two groups.

Subjects in the Ferguson and Kelley experiment had an extensive period of familiarization and personal investment in the outcome, which could have influenced their preference for own-group products. A clearer demonstration of in-group bias is obtained when subjects are asked to evaluate qualities associated with their own and other groups in the absence of any interaction or personal influence on the qualities being rated, as was the case in an experiment by Doise et al. (1972). Subjects in that experiment were divided into "X-type" and "Y-type" groups and were told that the division was based on photograph preferences (although group assignment was actually determined randomly). In the control condition of the experiment, subjects were led to anticipate no further interaction with members of either group, but were asked to describe the other members of their own group and the members of the other group on a series of 19 evaluative trait ratings. Despite the minimal basis for distinction between the two groups, a significant difference in mean favorableness of ratings was obtained in the direction of more positive ratings of members of the subjects' own group. However, in an earlier experiment, Rabbie and Horwitz (1969) found that in a control condition in which subjects were arbitrarily divided into groups labeled *blue* or *green* (with no rationale or further interaction), there were no

significant differences in evaluative trait ratings of individuals in the subjects' own group as opposed to individuals in the other group.

It appears, then, that there are lower limits to the effects of grouping on interpersonal perception but that in-group bias does occur in the absence of explicit competitive interdependence between groups. The absence of implicit competitive orientation in most of these studies, however, is difficult to establish. Indeed, Rabbie and Wilkens (1971) reported that their attempt to create coacting groups under *no-competition* instructional conditions resulted in ratings of perceived competitiveness that were equal to those obtained under explicit *competition* instructions. As Turner (1975) has suggested, the effect of categorization into groups may be mediated by an inherent competition for "positive social identity." Relative to the earlier view of the role of competition in intergroup attitudes, however, this hypothesis reverses the causal ordering in that competition is generated by the differentiation between groups rather than vice versa.

The generation of competitive orientation as a function of in-group–out-group distinctions, in the absence of any functional conflict of interests, is perhaps best illustrated with the paradigm originated by Tajfel and his colleagues for studying intergroup behavior (Tajfel, 1970; Tajfel, Billig, Bundy, & Flament, 1971). The research setting was designed to meet the following criteria for "minimal differentiation" (Tajfel et al., 1971, pp. 153–154): (a) no face-to-face interaction among subjects, within or between groups, (b) anonymity of group membership, (c) absence of any instrumental link between the basis for intergroup categorization and the response measure, and (d) a response measure involving real and significant choices but of no direct utilitarian value to the subject. Following these criteria, subjects in the Tajfel experiments are divided into two groups based supposedly on their responses to an irrelevant judgmental or preference test. After subjects are informed of their own group membership (but in the absence of any contact with or knowledge of other group members), they are given a choice task that involves allocating money between two other subjects in the same experiment. The identity of the other subjects is indicated only by an arbitrary identification number and a label specifying group membership, which can be varied to be the same as that of the subject or to indicate a member of the other group.

The types of choice matrices provided in the Tajfel experiments are illustrated in Table 1, for the case in which one target person is a member of the subject's own group and the other a member of the out-group. Within each matrix, each column represents an alternative allocation of points (worth some specified fractional amount of money) distributed between the two target persons, and the subject is to choose one of the alternatives as the distribution to be made. Matrices are constructed to represent a number of different possible distribution rules that could be applied, including equality (choosing the alternative that comes closest to giving each person the same number of points), maximizing joint outcome (choosing the alternative for which the total number of points is highest), or in-group favoritism (choosing the alternative that affords the in-group member more than the out-group member). For instance, Matrix A in Table 1 pits equality (choices at the middle of the series) against favoritism (choices toward the extreme right), whereas Matrix B varies favoritism (choices at the left), equality (midpoint choices), and joint outcomes (choices at the right).

Across a series of studies that used this allocation task (Billig & Tajfel, 1973; Tajfel & Billig, 1974; Tajfel et al., 1971), Tajfel and his colleagues have found that competitive choices favoring the in-group member tend to dominate over alternative available choice strategies. However, the matrices used in these studies have not been systematically varied to compare favoritism with all possible choice combinations. In particular, choice alternatives that maximize relative gain (i.e., the choice that maximizes the difference between in-group and out-group points in favor of the in-group member) have usually been confounded with alternatives that maximize absolute gain (i.e., the same choice maximizes the number of points that can be provided to the in-group member alone; cf. Matrices A and B in Table 1). Thus, the task structure

Table 1
Multiple-Choice Allocation Matrices[a]

Matrix	Payoffs for members of in-group and out-group													
A														
In-group	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Out-group	14	13	12	11	10	9	8	7	6	5	4	3	2	1
B														
In-group	19	18	17	16	15	14	13	12	11	10	9	8	7	
Out-group	1	3	5	7	9	11	13	15	17	19	21	23	25	
C														
In-group	7	8	9	10	11	12	13	14	15	16	17	18	19	
Out-group	1	3	5	7	9	11	13	15	17	19	21	23	25	

[a] Adapted from Tajfel, Billig, Bundy, and Flament (1971).

itself may have dictated a competitive strategy, in that gain for the in-group could be achieved only at cost to the out-group member. Only one matrix format (Matrix C in Table 1) has been used in which the alternative that maximizes the in-group member's outcome is different from the relative gain choice, and in this case the former is confounded with maximizing joint gain and maximizing the difference in favor of the out-group member.

To test the generality of preference for outcomes that maximize the competitive advantage to the in-group under forced-choice conditions, Brewer and Silver (1978) constructed a series of two-choice matrices to represent all possible pairings of the four alternative distribution rules of interest—equality, joint gain, relative gain, and absolute (in-group) gain (cf. MacCrimmon & Messick, 1976; McClintock, Messick, Kuhlman, & Campos, 1973). The matrices they used are reproduced in Table 2.[1] For each pair of two-choice matrices, the distribution rules that are confounded in the first matrix of the pair are opposed in the second matrix. Thus, assuming consistency of choice preferences across matrices, the pattern of choices for the two matrices in each pair combined discriminates perfectly among the four distribution rules, as indicated by the scoring key associated with each matrix pair in Table 2. Using this forced-choice format, Brewer and Silver found that a majority of subjects who had been divided into groups following an "aesthetic preference" test selected point distributions that maximized relative gain in favor of the in-group over choices that maximized absolute in-group gain or other alternatives. These results confirm that subjects treat in-group–out-group outcomes as competitively interdependent even when such an orientation is not required by the nature of available alternatives.

Sources of Variation in Bias

The research paradigms provided by Sherif et al. (1961) and by Tajfel et al. (1971) represent two extremes of the conditions of intergroup differentiation under which the occurrence of in-group bias may be studied. The Sherif field studies created a high degree of interaction and cooperative interdependence within groups combined with explicit competitive interdependence between groups, whereas the Tajfel laboratory studies involved minimal intragroup relationships and no predetermined functional interdependence between groups. Both types of research yield evidence of the presence of in-group favoritism, but results are not directly comparable for purposes of assessing variations in extent or intensity of such bias. Most of the experimental studies undertaken in this area since 1960 can be viewed as attempts to determine the contribution to in-group bias of various settings in

[1] In the task booklets actually used in the Brewer and Silver (1978) study, the matrices from these pairs were randomly intermixed in order of presentation.

between the extremes represented by the Sherif and Tajfel paradigms.

Table 3 provides a two-way classification of experimental studies published since 1960, in which at least one aspect of the conditions of intergroup differentiation has been systematically varied. Along with a classification of the major independent variables that have been manipulated, the studies listed in Table 3 are categorized according to which dependent variables were assessed of the three most widely used types of measures of in-group bias: (a) subjective ratings of individual group members or of the group membership as a whole on a series of evaluative trait scales, (b) ratings of the quality of group process (e.g., cohesiveness and cooperative atmosphere) or product, and (c) behavioral measures involving resource distribution decisions (e.g., Prisoner's Dilemma Game choices or the Tajfel allocation task). Studies that manipulated more than one independent variable or that included more than one dependent measure are multiply listed in Table 3.

Actual or Anticipated Competition

A number of the studies already reviewed indicated that a competitive reward structure is not a necessary precondition for obtaining significant in-group bias, but whether bias is increased by the presence of explicit conflict of interest between groups remains in question. The early laboratory studies of in-group bias, undertaken in the context of management training groups (Bass & Dunteman, 1963; Blake & Mouton, 1961), compared evaluations of the in-group obtained before and after introduction of a problem-solving competition against one or more other groups and reported consistent increases in positive self-evaluations during the intergroup competition phase.

In a more systematically controlled study of the effects of anticipated and actual competition, Rabbie and Wilkens (1971) divided subjects arbitrarily into pairs of three-person groups and then led both groups either to expect no further interaction or to expect to engage in an interactive task either in competition with the other group or independent of the other group (no competition). (Subjects who initially anticipated no interaction were later placed in the *competition* or *no-competition* conditions.) Ratings of own- and other-group members on six evaluative traits were then obtained from each subject prior to the interaction phase of the experiment and were again obtained from subjects in the competition and no-competition conditions after the interactive task. Before interaction, subjects who anticipated the task gave ratings of in-group members that were significantly higher than those obtained from subjects not expecting interaction, but subjects in all conditions showed equally significant bias in the difference between in-group and out-group ratings. Following interaction, the degree of bias in favor of own-group members increased significantly for subjects in both the competition and no-competition settings (although ratings of group products showed no signifi-

Table 2
Forced-Choice Allocation Matrices[a]

Matrix pair	Payoff 0	Payoff 1	Scoring key
A1			
In-group member	7	8	0, 0: Equality
Out-group member	9	4	0, 1: Joint gain
A2			1, 0: Relative gain
In-group member	7	8	1, 1: In-group gain
Out-group member	9	12	
B1			
In-group member	6	7	0, 0: Joint gain
Out-group member	8	3	0, 1: Equality
B2			1, 0: In-group gain
In-group member	6	5	1, 1: Relative gain
Out-group member	8	4	
C1			
In-group member	6	7	0, 0: Equality
Out-group member	4	10	0, 1: Relative gain
C2			1, 0: Joint gain
In-group member	6	7	1, 1: In-group gain
Out-group member	4	1	
D1			
In-group member	7	9	0, 0: Relative gain
Out-group member	5	12	0, 1: Equality
D2			1, 0: In-group gain
In-group member	7	6	1, 1: Joint gain
Out-group member	5	7	

[a] Adapted from Brewer and Silver (1978).

Table 3
Summary of Studies That Varied Conditions of Intergroup Differentiation

Experimental condition	Bias measure used: Evaluative trait rating	Bias measure used: Process/product evaluation	Bias measure used: Behavioral choices
Competition/ noncompetition	Brewer & Silver (1978) Doise et al. (1972) Goldman, Stockbauer, & McAuliffe (1977) Kahn & Ryen (1972) Rabbie & Wilkens (1971) Ryen & Kahn (1975) Worchel, Andreoli, & Folger (1977)	Bass & Dunteman (1963) Blake & Mouton (1961) Janssens & Nuttin (1976) Rabbie & Wilkens (1971) Rabbie, Benoist, Oosterbaan, & Visser (1974)	Brewer & Silver (1978) Doise et al. (1972)
Group outcome (success/failure)	Kahn & Ryen (1972) Rabbie & Horwitz (1969) Ryen & Kahn (1975) Wilson & Miller (1961)	Bass & Dunteman (1963) Blake & Mouton (1961) Worchel, Lind, & Kaufman (1975)	Branthwaite & Jones (1975)
Out-group similarity	Brewer & Silver (1978) Hensley & Duval (1976)		Allen & Wilder (1975) Billig & Tajfel (1973) Brewer & Silver (1978) Dion (1973) Wilson & Kayatani (1968)
Categorization salience	Dion & Earn (1975) Doise & Sinclair (1973) Gerard & Hoyt (1974) McKillip, Dimiceli, & Luebke (1977) Stephenson, Skinner, & Brotherton (1976)	Dustin & Davis (1970) Gerard & Hoyt (1974)	Billig & Tajfel (1973) Turner (1975)

cant in-group bias). Thus, the effect of actual intragroup interaction was to enhance favoritism toward in-group members, but equally so for competitive and independent groups. Similarly, Janssens and Nuttin (1976) found that members of interacting groups overestimated group successes more than did members of noninteracting groups but that groups who engaged in intergroup competition did not overestimate significantly more than groups who coacted independently. However, as was mentioned previously, the success of initial instructions for creating a noncompetitive task structure may be questionable, since a manipulation check in the Rabbie and Wilkens (1971) experiment revealed that *felt* competitiveness was equally high for members of the competition and no-competition groups.

A clearer manipulation of the structure of interdependence between groups is attained when conditions promoting intergroup competition are compared with conditions requiring cooperation between groups. One method for varying this feature of intergroup relations is through the use of instructional sets designed to induce competitive or cooperative orientation on the part of members of one group toward those of another group. Such an instructional manipulation was used in an experiment by Rabbie, Benoist, Oosterbaan, and Visser (1974) in which three-person groups were instructed to role play a team of union negotiators preparing for a meeting with a management team. After a 10-minute discussion period within the union group, subjects were asked to make ratings of the atmosphere in their own group and of their expectations regarding interactions with the management group. No significant differences in ratings of in-group cohesion or satisfaction were obtained between subjects in the competitive and cooperative orientation conditions, but members of competitive groups did report anticipating greater hostility toward the out-group than did members of groups in the cooperative condition.

Other methods of varying cooperation–competition involve direct manipulation of the structure of the intergroup task. One experiment reported by Kahn and Ryen (1972) used a simulated game setting in which three-person teams anticipated either cooperative or competitive interaction with another team. Before any actual interaction, subjects made ratings of their own team members and of out-group members on 11 evaluative semantic differential scales. A significant difference in ratings in favor of in-group members was obtained from subjects in the cooperative condition, but the size of this difference was significantly greater for subjects in the competitive condition. Such enhancement of in-group bias as a function of intergroup competition has not, however, proved reliable across research studies. Doise et al. (1972) divided subjects into two groups based, supposedly, on preference for photographs, and then led subjects in the experimental groups to anticipate a Tajfel-type money allocation task involving members of both groups. Instructions for allocation were varied to emphasize competitive own-gain maximization (outcomes to be distributed differentially between the groups) or joint-gain maximization (total outcomes to be divided equally between the two groups). Before the allocation task began, subjects evaluated in-group and out-group members on 19 trait scales. Mean ratings from subjects in both the cooperative and competitive conditions showed an in-group bias significantly greater than that obtained from control groups (who anticipated no future task), but the bias for competitive teams was not significantly different from that for cooperative teams (even though behavior afterwards, in the allocation task itself, did differ in a direction consistent with instructions).

One possible explanation of the difference in findings obtained by Kahn and Ryen and by Doise et al. is that the salience of the cooperation–competition manipulation may be highly variable when its impact is assessed prior to actual intergroup interaction. Brewer and Silver (1978) obtained trait ratings of in-group and out-group members from some subjects before they completed an allocation task and from other subjects after the allocations had been completed. Instructions for the allocation task were varied to generate a cooperative intergroup reward structure (achievement determined by adding each in-group member's points to an out-group mem-

ber's points), a competitive reward structure (achievement determined by the difference between points allocated to the in-group member and those allocated to the out-group member), or a reward structure based on total points allocated to the in-group member, regardless of out-group gains. As in the Doise et al. study, performance on the allocation task was significantly affected by these different instruction conditions. Subjects in both the independent and the competitive intergroup conditions predominantly made choices that maximized relative gain in favor of the in-group member, but subjects working under cooperative instructions made fewer relative-gain choices and more choices that maximized joint gain or equality between in-group and out-group member outcomes. However, trait ratings were significantly biased in favor of own-group members by subjects in all conditions, regardless of intergroup reward structure or of whether ratings were obtained before or after the behavioral measure.

Contrary to the Brewer and Silver (1978) findings, Worchel, Andreoli, and Folger (1977) found that intergroup competition significantly increased differential attraction between in-group and out-group in comparison with cooperative or independent intergroup settings. In the first phase of the Worchel et al. experiment, subjects were divided arbitrarily into two groups of four to six persons each. Members of each group were to work together on a joint product that would later be evaluated either in competition with, independently of, or in combination with the product of the other group. After an initial period of interaction, subjects were asked to rate their liking for each of the members of their own and the other group. Only small differences between conditions were obtained for mean attraction ratings of in-group members, but liking for out-group members was significantly higher in the cooperative setting than in the independent setting, and out-group attraction in competitive groups was significantly lower than for either cooperation or independence. On the other hand, Ryen and Kahn (1975) found that competitive interaction with an out-group increased in-group bias, over that obtained in cooperative conditions, by increasing evaluative ratings of in-group members but having no significant effect on out-group ratings.

Since competition is sometimes found to enhance in-group bias effects and sometimes found to have no additional impact,[2] it may be that intergroup competition does not affect intergroup attitudes directly, but only when confounded with other aspects of group differentiation. In other words, the presence of explicit competition may serve to clarify the distinction between in-group and out-group under conditions in which the differentiation would otherwise be ambiguous. The role of intergroup competition in clarifying in-group boundaries can be illustrated with an experiment by Goldman, Stockbauer, and McAuliffe (1977) in which effects of cooperation and competition were compared for both intergroup and intragroup reward structure. In their experiment, two-person teams interacted on a joint task in which achievement outcomes *within* teams were either cooperatively or competitively interdependent, while performance *between* groups was assessed either jointly (cooperatively) or competitively. Evaluative ratings of own-team members were significantly higher under conditions of intragroup cooperation than under intragroup competition, regardless of intergroup reward structure. However, the effects of intragroup competition on task performance were significantly less in the presence of intergroup competition than in the presence of intergroup cooperation. It is very likely that in the latter condition there was no perceptual differentiation between the second member of one's own team and the members of the other team. Only in the presence of a negative correlation between a subject's own final outcomes and those of the other team could such a differentiation be made, which in turn moderated the effects of the competitive task structure within teams. However, Rabbie and Wilkens (1971) reported that intragroup

---

[2] It should be noted that for those studies reporting no significant differences in bias between competitive and cooperative conditions, it is not a matter of results falling just short of statistical significance but rather that the in-group bias obtained is virtually the same under the two conditions.

status differentiation among members of three-person groups increased under conditions of intergroup competition, as compared with cooperative intergroup settings.

Group Outcomes: Success–Failure

One factor that is inherently confounded with the presence of explicit competitive interdependence between groups is that of differential shared fate; that is, under conditions of a competitive reward structure, members of a group share (or anticipate sharing) a common outcome that is distinct from the outcome shared by members of the other group. Such a co-occurrence of group boundaries and common fate is one of the criteria for perceived "entitivity" of social groupings discussed by Campbell (1958). The importance of shared outcomes as a determinant of in-group bias was empirically verified by Rabbie and Horwitz (1969). In that study, arbitrary classification of subjects into two groups labeled *blue* and *green* alone produced no significant in-group bias. However, when the experimenter introduced a chance allocation rule whereby one group won a prize and the other did not, subjects showed a significant bias in evaluative trait ratings (made after the award had been announced) in favor of their own group, regardless of whether their group had won or lost or of what allocation rule had been applied.

Most of the studies of shared fate as a determinant of in-group bias have focused on the effect of group achievement—success versus failure—in a competitive setting. The early management studies by Blake and Mouton (1961) and by Bass and Dunteman (1963) included ratings taken at the end of the intergroup competition phase, after the winning and losing teams had been announced. In both cases, the self-ratings of the winning groups remained significantly inflated (as they were during competition), but the losing groups' self-ratings dropped, at least temporarily. Similarly, Wilson and Miller (1961) found that when win–loss outcomes were experimentally manipulated, evaluative ratings of teammates and out-group members were affected. In comparison with ratings made prior to competition, subjects on winning teams showed a significant increase in bias in favor of their own team members, whereas subjects on losing teams showed a smaller difference in ratings of in-group and out-group members in favor of the winning out-group. Ryen and Kahn (1975) also found that winning under competitive conditions significantly enhanced in-group bias in evaluative trait ratings but that feedback indicating one's own group had lost reduced perceived in-group–out-group differences to nonsignificance.

One experiment reported by Kahn and Ryen (1972) extended the range of win–loss outcomes studied by having groups of subjects engage in a series of simulated football plays and then giving each group feedback information that they had won either 100%, 50%, or 0% of the plays in comparison with another team. After this feedback, subjects made ratings of in-group and out-group team members on 11 evaluative scales. Mean in-group ratings increased (and, to a lesser extent, out-group ratings decreased) as a function of the percentage of in-group success. The resulting differences between mean ratings of in-group and out-group were not significantly different from zero for groups with no wins (mean in-group–out-group difference $= -1.5$), were significantly biased in favor of one's own group for those with 50% wins (mean difference $= 8.3$), and were significantly more biased for those with 100% success (mean difference $= 14.5$).

In the preceding experiment, in-group bias occurred even when outcomes for the in-group were the same as those obtained by the out-group (50% wins), as long as the in-group attained some degree of success. In a second experiment, Kahn and Ryen (1972) tested whether differentiation between in-group and out-group outcomes was an important factor in in-group bias when competitive interdependence between groups was removed. Groups of three subjects worked independently on selected IQ test items, and then each group was given feedback indicating whether their performance resulted in a high proportion of successes or a low proportion of successes (high failure) and also whether

the other group's performance was high or low, with in-group and out-group results manipulated independently. After this feedback was provided, evaluative ratings were obtained of in-group and out-group members. Under these noncompetitive conditions, subjects in all conditions showed a bias in favor of their own group, but the degree of bias was significantly enhanced only when in-group success was combined with out-group failure.

Across these studies involving group performance outcomes there appears to be a consistent tendency for subjects to exaggerate the difference between in-group and out-group qualities when the in-group does well in comparison with the other group but to reduce the perceived difference when in-group and out-group performed the same or when the in-group does more poorly. Such a pattern serves to maximize favorable comparisons and to minimize unfavorable ones and may be typical of responses to single, or one-time, intergroup comparisons. Responses to success and failure may change, however, if interactions are extended in time and further comparisons between in-group and out-group are anticipated in the future. Worchel, Lind, and Kaufman (1975) found that anticipation of further competition interacted with outcome feedback in determining relative evaluation of in-group and out-group products. Members of winning groups overevaluated their group product less when they expected competition between the groups to continue than when they expected it to discontinue, whereas members of losing groups devaluated their group product more under discontinuing than under continuing conditions. Worchel et al. interpreted these results in the context of ongoing competition to avoid "complacency" on the part of winning groups and to avoid "giving up" on the part of losing groups.

Continued failure or deprivation of the in-group relative to a particular out-group across a long period of time may lead to compensatory overevaluation in favor of the in-group wherever possible (LeVine & Campbell, 1972). Branthwaite and Jones (1975) looked at the effect of long-standing status differentials between ethnic groups on allocations in the Tajfel choice task. When subjects were divided into groups according to ethnic identity (Welsh–English), members of the minority group made more choices that maximized the difference between in-group and out-group member outcomes than did members of the majority group (who tended to make more choices dictated by equality or joint-gain maximization). Similar findings were obtained by Gerard and Hoyt (1974) with experimentally created groups and a different measure of bias. Subjects in their study were classified as members of a group of 2, 5, or 8 subjects, out of a total of 10 subjects participating in a session. Each subject was then asked to make evaluative ratings of essays supposedly written by two other participants in the experiment—one identified by an identification number of someone in the subject's own group and one identified by an identification number from the out-group. Ratings of the content of the essays produced no in-group–out-group differentiation, but evaluations of the writers resulted in some differences. Subjects classified into groups of five and eight showed no significant bias (in fact, there was some tendency in favor of the out-group member), whereas subjects in the minority group of two showed a significant positive bias in favor of their in-group member.

Results from both of these studies suggest that minority group status makes in-group membership more salient than does membership in a majority group. A similar heightening of awareness of group identity may occur for groups exposed to repeated failure or loss, particularly when membership in such a group is unalterable. Whether repeated failure ultimately generates greater in-group–out-group differentiation than does repeated success has yet to be experimentally demonstrated. However, in a survey of intergroup perceptions among ethnic groups in East Africa, Brewer and Campbell (1976) found those groups rated lowest on the socioeconomic status index to be higher in ethnocentric self-regard than those groups with the highest socioeconomic status ratings. This effect may also be related to repeated findings

of "reverse discrimination" on the part of members of high-status majority groups in dealings with individual members of minority groups (Dutton, 1976).

Intergroup Similarity

Although it has been established that evaluative bias occurs only in the presence of some meaningful distinction between groups (Rabbie & Horwitz, 1969), the minimal differentiation required allows room for considerable variation in implied or explicit similarity between members of the in-group and out-group. A number of studies have examined the effect on in-group bias of variations in degree of similarity among in-group members, or of dissimilarity between in-group and out-group, on such dimensions as cultural, personality, or attitudinal characteristics.

Wilson and Kayatani (1968) divided subjects into two-person teams, with each team composed of members of the same racial group (Japanese or Caucasian). Each team then played a modified Prisoner's Dilemma Game with another team of the same or different race. The game choice results were uniform across both types of out-groups—choices made in the intergroup setting averaged only 43% cooperative, whereas choices made within each group averaged 84% cooperative. Similarly, postgame evaluative trait ratings showed a significant in-group bias regardless of whether the out-group was of the same or different race.

A more recent study by Dion (1973) also looked at the effect of similarity on intergroup versus intragroup Prisoner's Dilemma Game behavior. The experimental manipulation in this study, however, varied intragroup rather than intergroup similarity. One half of the dyads in the experiment were told that the members had closely matched personality profiles, whereas the remaining pairs were told they had discrepant profiles. All teams then played a Prisoner's Dilemma Game (with two experimental confederates serving as the out-group team) and also rated both in-group and out-group members on 16 evaluative traits. Both high- and low-similar dyads exhibited the same intergroup game behavior (averaging 30% cooperative choices), but the high-similar pairs exhibited significantly more in-group cooperation (59%) than did the low-similar pairs (36%). The same pattern of results was obtained for the evaluative ratings: Out-group ratings were essentially the same for all groups, while in-group ratings were significantly higher for members of the high-similar dyads.

Billig and Tajfel (1973) compared intergroup differentiation based on explicit similarity with categorization based on no similarity principle. In their experiment, intergroup similarity and categorization were manipulated independently. Similarity was varied by dividing half the groups according to supposed preferences in painting styles (Klee vs. Kandinsky) and by dividing the remaining groups randomly into groups labeled *X* or *W*. Categorization was varied by including group label as part of subject identification during the allocation task for some subjects and omitting group labels for others. Results from the allocation task showed significant in-group favoritism in the categorization conditions and no significant favoritism in the noncategorization conditions, regardless of similarity. Brewer and Silver (1978) also found significant in-group bias on both allocation-task decisions and evaluative ratings regardless of whether groups had been formed on the basis of distinct preferences or had been formed on the basis of a random split after being explicitly told that all subjects were similar on the preference task.

The similarity manipulation in the Billig and Tajfel and the Brewer and Silver studies involved both intragroup similarity and intergroup dissimilarity. In an experiment by Allen and Wilder (1975) these two facets were varied independently in a 2×2 design. With painting style preference as the ostensible basis of categorization into groups, subjects were provided with further information indicating the percentage (high or low) of responses to an attitude questionnaire that were similar to their own responses for other members of their own group and for members of the other group. Subjects then

made choices on the Tajfel allocation task on behalf of an in-group member and an out-group member. Subjects in all experimental conditions showed some degree of in-group favoritism in allocation decisions. High in-group similarity produced significantly more bias than did in-group dissimilarity, but similarity–dissimilarity of the out-group had no effect on degree of in-group bias.

Results from all of these studies are consistent in indicating that explicit dissimilarity within the in-group reduces in-group bias but that information on similarity between the subject and out-group members makes no difference. However, it may be that perceived similarity within the in-group and perceived dissimilarity from the out-group are highly interdependent, as suggested by the results of an experiment by Hensley and Duval (1976). In this study, information on the opinions held by 10 subjects in a discussion group was presented graphically in such a way that each subject's own opinion was depicted within a cluster of seven other subjects' opinions, with the distance between the subject's opinion and these seven held constant. The positioning of the opinions associated with the remaining 2 subjects was varied across five levels of distance from this majority cluster. Following this presentation, each subject made ratings of the other 9 subjects in the session on perceived similarity to self and on liking. The results for perceived similarity ratings revealed an assimilation–contrast effect: The greater the distance between the minority (out-group) and the majority (in-group), the greater the perceived similarity within the subject's own group. A parallel effect was obtained on the ratings of liking for members of the majority and minority groups.

Salience of Categorization

Since the grouping of subjects into majority and minority clusters in the Hensley and Duval study was not explicitly labeled, it is likely that the effect of increasing the visual distance between the two clusters was to increase the probability that the subject would perceive a boundary between the two groups. In fact, in the three conditions in which the distance between clusters was great enough to insure the perception of distinct groupings, the ratings of perceived similarity and of liking for in-group as opposed to out-group members were essentially the same, the only significant differences occurring between these three conditions and the two conditions involving lesser distances. Thus, the effect attributed to out-group dissimilarity may have been due to the differential salience of the in-group–out-group distinction. Other research in which the salience of categorization has been manipulated either directly (e.g., Billig & Tajfel, 1973) or indirectly (e.g., Gerard & Hoyt, 1974) confirms the importance of this factor in eliciting in-group bias.

Results of several studies indicate that the same differences among individuals may or may not lead to bias depending on whether a basis for grouping has been made salient. For instance, in a study by Stephenson, Skinner, and Brotherton (1976), secondary school students were assessed on their attitudes toward raising the age for compulsory schooling. Experimental sessions were composed of four students in favor of and four against raising the age. At the beginning of the session subjects were given information about the distribution of attitudes among the eight participants and then were asked to rate each of the participants on five evaluative trait scales. Ratings were made both before and after subjects were divided into four-person groups (based on initial attitudes) for participation in an intergroup negotiation task. Prior to the division into labeled groups, ratings showed no in-group bias, but following the group task, ratings changed significantly in favor of the in-group.

Turner (1975) also reported a complex interaction between participation in an intergroup task and in-group favoritism. Subjects in his study were divided into two groups and then made two sets of choices on a Tajfel allocation task—once making choices on behalf of two other subjects (one of whom was an in-group member and one an out-group member) and once making choices on behalf

of self and one other subject (who was either an in-group member or an out-group member). For subjects who made self–other choices first, favoritism toward self was moderately high, regardless of whether the other was an in-group member or an out-group member. However, for subjects who made in-group–out-group member choices before making self–other choices, self-favoritism was significantly higher when the other was an out-group member than when the other was an in-group member. Thus, prior participation in a task that made the intergroup distinction salient enhanced the differentiation between self and out-group member, but reduced differentiation between self and in-group member.

The mere presence of more than one member of a distinct social group apparently increases the salience of grouping and associated biases. Dustin and Davis (1970) observed the effects of competition between two groups of three subjects when the competitive interaction took place on an individual (1:1) basis or on a group (3:3) basis. Following group competition, product ratings were significantly biased in favor of subjects' in-group output, but no own-group bias was obtained for product ratings from subjects whose groups interacted on an individual competition basis. Similar effects have been obtained for biases associated with nonexperimental social groups. Doise and Sinclair (1973) studied the effect of reference group salience on accentuation of stereotypes associated with *collegians* (male secondary school students) and *apprentis* (vocational trainees). Members of both groups were brought together in either a 1:1 or a 2:2 encounter and, following a short discussion period, were asked to make trait ratings of the respective groups. In the 2:2 condition collegians gave ratings significantly more biased in favor of their own group than they did in the 1:1 condition, whereas apprentis showed less derogation of their own group relative to the higher status out-group in the 2:2 than in the 1:1 condition. Similar effects of the presence of multiple members of both groups have been obtained for accentuation of stereotypes based on sex (McKillip, Dimiceli, & Luebke, 1977) and on ethnic identity (Dion & Earn, 1975).

Summary

The interpretation of results from all of the experimental studies reviewed in this section (entitled Sources of Variation in Bias) has been consistent with the following general conclusion: Any of the situational factors found to be associated with enhancement of in-group bias can be subsumed under the effect of the salience of the distinction between in-group and out-group. Factors such as interdependence, intergroup similarity, and shared fate all affect the probability that a respondent will be aware of a relevant basis for categorization into groups, which in turn determines the amount of in-group bias that is evidenced. Once a particular categorization has become salient, however, the degree of bias obtained is fairly constant despite further variations in out-group similarity (e.g., Allen & Wilder, 1975) or in opportunity for cooperative interaction (e.g., Brewer & Silver, 1978; Worchel et al., 1977).

Locus of Bias: Which Dimensions?

Though the argument has been made that in-group bias is related in an all-or-nothing manner to category salience, the bias associated with any particular basis for categorization into in-group and out-group may not be constant across all response dimensions. There are a number of sources of evidence for specificity of effects, or "selective bias" (Wilson, Chun, & Kayatani, 1965). A series of studies by Wilson and his associates (e.g., Wilson et al., 1965; Wilson & Kayatani, 1968) indicate that following intergroup competition within a Prisoner's Dilemma Game format evaluative bias is most pronounced on game-relevant motive traits (e.g., cooperative, fair, and kind), whereas in-group bias is less pronounced on sociometric (e.g., likeable) or ability traits (e.g., capable and intelligent) and least evident on general personality dispositions (e.g., neurotic and

anxious). Dion (1973) also found that in-group bias after participation in an intergroup Prisoner's Dilemma Game was greatest on the dimension of trust, and Brewer and Silver (1978) obtained the most bias on ratings of trustworthiness, friendliness, and cooperativeness, even after respondents had engaged in a cooperative intergroup allocation task. The latter study also found a nonsignificant correlation ($r = .14$) between in-group favoritism on the allocation task and in-group bias on evaluative trait ratings. Similarly, Ryen and Kahn (1975) obtained no significant correlation between evaluative in-group bias and intergroup distancing, as evidenced in seating behavior, and Worchel et al. (1975) reported a low correlation between liking for the in-group and relative evaluation of in-group versus out-group products. Going outside of the laboratory, Brewer and Campbell (1976) found in a large-scale survey of intergroup attitudes that the psychological distance reported between a respondent's in-group and a particular out-group varied depending on whether the response measure dealt with affective relations (e.g., social distance), evaluation, or respect.

The finding that bias depends on some interaction between the categorization variable and the response dimension on which bias is assessed is consistent with a cognitive interpretation of intergroup bias (Tajfel, 1959; 1969). The comparison with general theories of cognitive processing is best illustrated by Tversky's (1977) feature-matching model of similarity judgments. In Tversky's model, the perceived similarity between two objects is a function of some linear combination of their common and distinctive features; but the weight assigned to any particular feature or set of features may vary depending on the context or the nature of the judgment task. As a result, the same two objects may be judged to be highly similar within one frame of reference and highly distinct within another. The determinants suggested by Tversky for this lability of perceived similarity are relevant to the judgment of objects external to the respondent. It may be that when identification with oneself is a salient feature of one of the objects to be judged, motivational factors enter into the selection of features to be attended to (cf. Christian, Gadfield, Giles, & Taylor, 1976).

The interdependence of perceptual and motivational factors is highlighted by some interesting parallels between the research literature on perceptual accentuation in social judgment (Eiser & Stroebe, 1972) and on in-group bias effects (Turner, 1975). In the judgment literature, *enhancement of contrast* occurs when the judged distance between stimuli that are members of different classes is exaggerated. The occurrence of this accentuation effect depends on the presence of at least some minimal correlation between the classification variable and variation among the stimuli on the property being judged (Campbell, 1956; Tajfel, 1959; Tajfel & Wilkes, 1963). However, when the stimuli are social objects, toward which the judge has differential orientations, a second condition has been found to be necessary for enhancement of contrast, namely, that the judge's own position be located on what the rater perceives to be the positive side of the dimension of judgment (Eiser, 1975; Eiser & Mower-White, 1974).

Analogous to the conditions associated with the enhancement of contrast effect, Turner's (1975) social comparison theory of in-group favoritism specifies two preconditions: (a) salience of some basis for distinction between in-group and out-group and (b) availability of "differentially valued actions relevant to the categorization" (p. 12). The presence of these conditions generates intergroup *social competition,* the aim of which is to take advantage of opportunities to maximize the relative advantage of the in-group over the out-group. In effect, then, in-group bias results from a motivated search to represent the differences between groups along dimensions that favor the in-group. If outcomes favoring the in-group are not available, the distinction between them will be minimized rather than accentuated.

The presence of motivational influences can lead to important asymmetries in the ways in which the members of two social groups perceive the differences between them, such as those obtained for groups differing

in socioeconomic status (e.g., Branthwaite & Jones, 1975; Brewer & Campbell, 1976). Members of Group A may perceive a major difference between themselves and Group B along dimension *X*, whereas members of Group B may focus on the common features of A and B relevant to *X* and emphasize the distinctive features relevant to dimension *Y*. (Imagine, e.g., members of a winning team following a football contest and highlighting the differences between the teams in agility and skill, while members of the losing team regard ability differences as marginal, but emphasize differences between the teams in "how the game was played" with respect to fair play and sportsmanship.) Even differences on a single dimension can be represented in alternative ways that favor one group or the other (Campbell, 1967; Peabody, 1967; Vassiliou, Triandis, Vassiliou, & McGuire, 1972). For example, an objective difference between the customs of two groups with regard to the sharing of household items may be represented by one group as a distinction between generosity and selfishness, but may be defined by the other group in terms of responsibility–irresponsibility. Such differences can introduce considerable variation in assessments of in-group bias.

Although not all differences can be represented in a manner congruent with positive self-image, some characteristics of groups lend themselves to universal bias. Across experimental and field studies of the content of intergroup perceptions, the dimensions on which evaluative bias in favor of in-groups occurs most reliably are those associated with trustworthiness, honesty, or loyalty. All these are traits related to normative expectations that apply to intragroup—as opposed to intergroup—behavior. To the extent that norms prescribing preferential treatment for members of one's own group are characteristic of in-group formation, they generate a set of reciprocal stereotypes (Campbell, 1967) that any two groups might have of each other and with which each could legitimately place the in-group on the positive side of the scale (e.g., "we are loyal; they are clannish"; "we are honest and peaceful among ourselves; they are hostile and treacherous toward outsiders"). This reciprocal contrast is basic to the "mirror-image" phenomenon in international perception, as portrayed by Bronfenbrenner (1961).

Locus of Bias: In-Group or Out-Group?

The extensive literature on group cohesiveness indicates that factors such as similarity among group members (e.g., Anderson, 1975) and shared success (e.g., Blanchard, Adelman, & Cook, 1975) enhance attraction toward one's own group in the absence of comparison with any other groups. Since in-group bias research focuses on favoritism toward the in-group *relative* to an out-group, it is often ambiguous whether the comparison rests on enhancement of the in-group, devaluation of the out-group, or both. In many studies, particularly those dealing with evaluative biases, results were reported only in the form of net ratings or difference scores (e.g., Doise & Sinclair, 1973; Dustin & Davis, 1970; Ferguson & Kelley, 1964; Gerard & Hoyt, 1974; McKillip et al., 1977), thereby losing information as to whether variations in bias were a function of increases in in-group ratings or decreases in out-group ratings.

Among those studies that did report both in-group and out-group ratings separately, results are mixed as to the location of bias. Some studies that compared intergroup cooperation and competition reported no change in in-group attraction, but reported a decrease in out-group ratings under competition conditions (Rabbie et al., 1974; Worchel et al., 1977). Other research indicates that variations in degree of bias are a function of both increased in-group and decreased out-group ratings (Hensley & Duval, 1976; Kahn & Ryen, 1972; Wilson et al., 1965). The majority of studies, however, indicate that increases in bias are associated with enhanced in-group evaluation, whereas out-group ratings remain relatively constant (Dion, 1973; Rabbie & Horwitz, 1969; Rabbie & Wilkens, 1971; Ryen & Kahn, 1975; Stephenson et al., 1976; Wilson & Miller, 1961; Worchel et al., 1975). The results in general, then, are consistent with the conclu-

sion that in-group bias rests on the perception that one's own group is better, although the out-group is not necessarily depreciated.

The above conclusion suggests that the effect of in-group–out-group categorization is one of differentiating the in-group from the out-group rather than of differentiating the out-group from the in-group, as the process is usually conceived. This means that the baseline should be conceptualized as a state in which the self is perceived as distinct from an undifferentiated group of others. The introduction of an in-group–out-group boundary is then associated with a realignment of perceptions wherein members of the in-group are perceived to be less differentiated from the self, while the distance between the self and out-group members remains unchanged. This conceptualization of the differentiation process is borne out by studies that modified the Prisoner's Dilemma Game for group play (e.g., Dion, 1973; Wilson & Kayatani, 1968). In terms of the high percentage of competitive choices, intergroup behavior in these games parallels closely the game behavior of individual players. It is the increased proportion of cooperative choices exhibited in intragroup decisions that deviates from typical interindividual play.

Reconceptualizing the process of intergroup differentiation tends to shift the focus of attention from the negative implications of out-group perceptions to the positive consequences of in-group formation. The critical role of in-group identity in the extension of interpersonal trust has already been alluded to. Another consequence of the reduced social distance between self and others that accompanies in-group formation is that outcomes to other group members, or to the group as whole, come to be perceived as one's own. Indeed, there is evidence that feedback regarding total group outcomes can have more impact on the individual than feedback on his or her own performance (e.g., Zander & Armstrong, 1972) and that expected and perceived success is higher at the group level than at the individual level (Janssens & Nuttin, 1976). Satisfaction and identification with group success tend to be high even when the individual's contribution to that success has been minimal (e.g., Kahn & Ryen, 1972) or nil (Cialdini et al., 1976).

The capacity of in-group identification to amplify feedback has important implications for the solution of that class of social problems characterized as *commons dilemmas* (Dawes, McTavish, & Shaklee, 1977; Hardin, 1968) or *social traps* (Platt, 1973). The essence of these problems is a "divergence between what people are individually motivated to do and what they might accomplish together" (Schelling, 1971, p. 68). The most critical social dilemmas derive from behaviors for which rewards outweigh small costs at the *individual* level (e.g., taking an extended shower) but that result in cumulative high costs at the *group* level (e.g., depletion of water supplies). The solution to such dilemmas requires that the collective outcome be real enough to the actor to overcome individualistic motivational dynamics (Messick, 1973; 1974). The reduced differentiation between one's own and other outcomes associated with in-group formation provides one mechanism for increasing the weight given to collective outcomes in individual decision making.

The idea of capitalizing on the social benefits of group identification raises concern about whether the positive consequences of in-group formation depend on the presence of a distinct out-group. To those who hold that the effects of social categorization are the result of intergroup social comparison (Turner, 1975; Tajfel, in press), the existence of an identifiable out-group is essential. Although groups may function in the absence of any other groups, the mere presence of an out-group is sufficient to significantly alter in-group processes (Billig, 1976). On the other hand, if one associates group identification with more general concepts of unit formation (Campbell, 1958; Heider, 1958), awareness of differentiated social groupings may be only one potential mechanism—however important—by which the self is included in a bounded social unit. Perhaps the salience of interdependence or common fate can be enhanced among any given set of individuals without reference to other subsets. If so, the focus of research on in-

group bias should be shifted from intergroup to intragroup contexts.

References

Allen, V. L., & Wilder, D. A. Categorization, belief similarity, and intergroup discrimination. *Journal of Personality and Social Psychology,* 1975, *32,* 971–977.

Anderson, A. B. Combined effect of interpersonal attraction and goal-path clarity on the cohesiveness of task oriented groups. *Journal of Personality and Social Psychology,* 1975, *31,* 68–75.

Bass, B. M., & Dunteman, G. Biases in the evaluation of one's own group, its allies, and opponents. *Journal of Conflict Resolution,* 1963, *7,* 16–20.

Billig, M. *Social psychology and intergroup relations.* London: Academic Press, 1976.

Billig, M., & Tajfel, H. Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology,* 1973, *3,* 27–52.

Blake, R. R., & Mouton, J. S. Reactions to intergroup competition under win-lose conditions. *Management Science,* 1961, *7,* 420–435.

Blanchard, F. A., Adelman, L., & Cook, S. W. Effect of group success and failure upon interpersonal attraction in cooperating interracial groups. *Journal of Personality and Social Psychology,* 1975, *31,* 1020–1030.

Branthwaite, A., & Jones, J. E. Fairness and discrimination: English versus Welsh. *European Journal of Social Psychology,* 1975, *5,* 323–338.

Brewer, M. B., & Campbell, D. T. *Ethnocentrism and intergroup attitudes: East African evidence.* New York: Halstead Press, 1976.

Brewer, M. B., & Silver, M. Ingroup bias as a function of task characteristics. *European Journal of Social Psychology,* 1978, *8,* 393–400.

Bronfenbrenner, U. The mirror-image in Soviet-American relations. *Journal of Social Issues,* 1961, *17,* 45–46.

Campbell, D. T. Enhancement of contrast as a composite habit. *Journal of Abnormal and Social Psychology,* 1956, *53,* 350–355.

Campbell, D. T. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science,* 1958, *3,* 14–25.

Campbell, D. T. Stereotypes and the perception of group differences. *American Psychologist,* 1967, *22,* 817–829.

Christian, J., Gadfield, N., Giles, H., & Taylor, D. The multidimensional and dynamic nature of ethnic identity. *International Journal of Psychology,* 1976, *11,* 281–291.

Cialdini, R. B., et al. Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology,* 1976, *34,* 366–375.

Dawes, R. M., McTavish, J., & Shaklee, H. Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology,* 1977, *35,* 1–11.

Dion, K. L. Cohesiveness as a determinant of ingroup–outgroup bias. *Journal of Personality and Social Psychology,* 1973, *28,* 163–171.

Dion, K. L., & Earn, B. M. The phenomenology of being a target of prejudice. *Journal of Personality and Social Psychology,* 1975, *32,* 944–950.

Doise, W., et al. An experimental investigation into the formation of intergroup representations. *European Journal of Social Psychology,* 1972, *2,* 202–204.

Doise, W., & Sinclair, A. The categorization process in intergroup relations. *European Journal of Social Psychology,* 1973, *3,* 145–157.

Dustin, D. A., & Davis, H. P. Evaluative bias in group and individual competition. *Journal of Social Psychology,* 1970, *80,* 103–108.

Dutton, D. G. Tokenism, reverse discrimination, and egalitarianism in interracial behavior. *Journal of Social Issues,* 1976, *32,* 93–107.

Eiser, J. R. Attitudes and the use of evaluative language: A two-way process. *Journal for the Theory of Social Behaviour,* 1975, *5,* 235–248.

Eiser, J. R., & Mower-White, C. J. Evaluative consistency and social judgment. *Journal of Personality and Social Psychology,* 1974, *30,* 349–359.

Eiser, J. R., & Stroebe, W. *Categorization and social judgement.* London: Academic Press, 1972.

Ferguson, C. K., & Kelley, H. H. Significant factors in overevaluation of own-group's product. *Journal of Abnormal and Social Psychology,* 1964, *69,* 223–228.

Gerard, H. B., & Hoyt, M. F. Distinctiveness of social categorization and attitude toward ingroup members. *Journal of Personality and Social Psychology,* 1974, *29,* 836–842.

Goldman, M., Stockbauer, J. W., & McAuliffe, T. G. Intergroup and intragroup competition and cooperation. *Journal of Experimental Social Psychology,* 1977, *13,* 81–88.

Hamilton, D. L. Cognitive biases in the perception of social groups. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior.* Hillsdale, N.J.: Erlbaum, 1976.

Hardin, G. The tragedy of the commons. *Science,* 1968, *162,* 1234–1248.

Heider, F. *The psychology of interpersonal relations.* New York: Wiley, 1958.

Hensley, V., & Duval, S. Some perceptual determinants of perceived similarity, liking, and correctness. *Journal of Personality and Social Psychology,* 1976, *34,* 159–168.

Janssens, L., & Nuttin, J. R. Frequency perception of individual and group successes as a function of competition, coaction, and isolation. *Journal of Personality and Social Psychology,* 1976, *34,* 830–836.

Kahn, A., & Ryen, A. H. Factors influencing the bias towards one's own group. *International Journal of Group Tensions,* 1972, *2,* 33–50.

LeVine, R. A., & Campbell, D. T. *Ethnocentrism:*

Theories of conflict, ethnic attitudes and group behavior. New York: Wiley, 1972.

MacCrimmon, K. R., & Messick, D. M. A framework for social motives. *Behavioral Science*, 1976, *21*, 86–100.

McClintock, C. G., Messick, D. M., Kuhlman, D. M., & Campos, F. T. Motivational bases of choice in three-choice decomposed games. *Journal of Experimental Social Psychology*, 1973, *9*, 572–590.

McKillip, J., Dimiceli, A. J., & Luebke, J. Group salience and stereotyping. *Social Behavior and Personality*, 1977, *5*, 81–85.

Messick, D. M. To join or not to join: An approach to the unionization decision. *Organizational Behavior and Human Performance*, 1973, *10*, 145–156.

Messick, D. M. When a little "group interest" goes a long way. *Organizational Behavior and Human Performance*, 1974, *12*, 331–334.

Peabody, D. Trait inferences: Evaluative and descriptive aspects. *Journal of Personality and Social Psychology Monograph*, 1967, *7*(4, Pt. 2).

Platt, J. Social traps. *American Psychologist*, 1973, *28*, 641–651.

Rabbie, J. M., Benoist, F., Oosterbaan, H., & Visser, L. Differential power and effects of expected competitive and cooperative intergroup interaction on intragroup and outgroup attitudes. *Journal of Personality and Social Psychology*, 1974, *30*, 46–56.

Rabbie, J. M., & Horwitz, M. Arousal of ingroup-outgroup bias by a chance win or loss. *Journal of Personality and Social Psychology*, 1969, *13*, 269–277.

Rabbie, J. M., & Wilkens, G. Intergroup competition and its effect on intragroup and intergroup relations. *European Journal of Social Psychology*, 1971, *1*, 215–234.

Ryen, A. H., & Kahn, A. Effects of intergroup orientation on group attitudes and proxemic behavior. *Journal of Personality and Social Psychology*, 1975, *31*, 302–310.

Schelling, T. On the ecology of micromotives. *The Public Interest*, 1971, *25*, 59–98.

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. *Intergroup conflict and cooperation: The Robbers Cave experiment*. Norman, Oklahoma.: University Book Exchange, 1961.

Stephenson, G. M., Skinner, M., & Brotherton, C. J. Group participation and intergroup relations: An experimental study of negotiation groups. *European Journal of Social Psychology*, 1976, *6*, 51–70.

Sumner, W. G. *Folkways*. Boston: Ginn, 1906.

Tajfel, H. Quantitative judgement in social perception. *British Journal of Psychology*, 1959, *50*, 16–29.

Tajfel, H. Cognitive aspects of prejudice. *Journal of Social Issues*, 1969, *25*, 79–97.

Tajfel, H. Experiments in intergroup discrimination. *Scientific American*, 1970, *223*(2), 96–102.

Tajfel, H. (Ed.). *Differentiation between social groups: Studies in the social psychology of intergroup relations*. London: Academic Press, in press.

Tajfel, H., & Billig, M. Familiarity and categorization in intergroup behavior. *Journal of Experimental Social Psychology*, 1974, *10*, 159–170.

Tajfel, H., Billig, M., Bundy, R., & Flament, C. Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1971, *1*, 149–178.

Tajfel, H., & Wilkes, A. L. Classification and quantitative judgement. *British Journal of Psychology*, 1963, *54*, 101–114.

Turner, J. C. Social comparison and social identity: Some prospects for group behavior. *European Journal of Social Psychology*, 1975, *5*, 5–34.

Tversky, A. Features of similarity. *Psychological Review*, 1977, *84*, 327–352.

Vassiliou, V., Triandis, H. C., Vassiliou, G., & McGuire, H. Interpersonal contact and stereotyping. In H. C. Triandis, *The analysis of subjective culture*. New York: Wiley, 1972.

Wilson, W., Chun, N., & Kayatani, M. Projection, attraction, and strategy choices in intergroup competition. *Journal of Personality and Social Psychology*, 1965, *2*, 432–435.

Wilson, W., & Kayatani, M. Intergroup attitudes and strategies in games between opponents of the same or of a different race. *Journal of Personality and Social Psychology*, 1968, *9*, 24–30.

Wilson, W., & Miller, N. Shifts in evaluations of participants following intergroup competition. *Journal of Abnormal and Social Psychology*, 1961, *63*, 428–431.

Worchel, S., Andreoli, V. A., & Folger, R. Intergroup competition and intergroup attraction: The effect of previous interaction and outcome of combined effort. *Journal of Experimental Social Psychology*, 1977, *13*, 131–140.

Worchel, S., Lind, E. A., & Kaufman, K. H. Evaluations of group products as a function of expectations of group longevity, outcome of competition, and publicity of evaluations. *Journal of Personality and Social Psychology*, 1975, *31*, 1089–1097.

Zander, A., & Armstrong, W. Working for group pride in a slipper factory. *Journal of Applied Social Psychology*, 1972, *2*, 293–307.

Received November 17, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 325–334

A Simplex Process Model for Describing Differences Between Cross-Lagged Correlations

Lloyd G. Humphreys and Charles K. Parsons
University of Illinois at Urbana-Champaign

A model that is based on the use of the diagonal method of factoring to describe a simplex process and that explains differences between cross-lagged correlations is presented. When the model is applied to the data of Atkin et al., which showed that Listening tapped most directly the causes of later intellectual development, quite good fits are obtained for both 2- and 4-year lags. The best fit, however, is based on an estimate of the effects of a 3-year lag between changes in the rank order of individual differences on the Listening test and the same changes on the intellectual composite. It is shown that accurate knowledge of reliabilities and specificities of the measures is necessary for the interpretation of cross-lagged differences but that stationarity per se is not an essential assumption.

The first reaction of most behavioral scientists to a discussion of cross-lagged panel correlation methodology is to question the inference of causality from a difference in the size of correlations. It does not matter how large or how statistically significant the difference may be. Discussions of this methodology, such as those by Campbell and Stanley (1963) and by Kenny (1975), have not provided a rationale that is sufficient to satisfy these skeptics. The purpose of this article is to describe a factor model that provides a new and plausible basis for a possible causal inference and to apply it to some published data.

Simplex Correlation Matrices

Since the cross-lagged methodology is used in the expectation that there will be change in the rank order of individual differences in the functions measured with the passage of time, one should also expect that the intercorrelations of each of the two or more measures over multiple occasions will show the simplex pattern (Guttman, 1955; Humphreys, 1960; Rozelle & Campbell, 1969). Guttman used the simplex matrix originally to describe the intercorrelations of several measures on a single measurement occasion when the several measures seemingly varied only in their degree of complexity. Humphreys extended the applicability of the simplex pattern to learning and maturational data in which one measure was obtained on repeated occasions. Rozelle and Campbell discussed the simplex matrix with respect to time series generally. Both Guttman and Humphreys argued against the use of multiple or common factor analysis as appropriate for such data.

In place of common factor analysis, Guttman recommended the use of diagonal factoring, which produces the same number of factors as there are variables in the correlation matrix. If the measures are ordered in increasing order of complexity, the first factor is based on the correlations that involve the simplest measure, and the nth is based on the residual correlations that involve the most complex measure. For learning data, if one assumes accretion of responses, the first factor is based on the correlations that involve the

This research was supported by National Institute of Mental Health Grant MH 23612-04. The authors wish to thank Robert Linn for assistance in the estimation of reliabilities and for a critical reading of the manuscript.

Requests for reprints should be sent to Lloyd G. Humphreys, 425 Psychology Building, University of Illinois, Champaign, Illinois 61820.

Table 1
A Simplex Matrix of Intercorrelations and Its Diagonal Factoring

Occasion	1	2	3	4
		Occasion		
1	1.000	.960	.922	.885
2	.960	1.000	.960	.922
3	.922	.960	1.000	.960
4	.885	.922	.960	1.000
		Factor		
1	1.000	.000	.000	.000
2	.960	.280	.000	.000
3	.922	.268	.280	.000
4	.885	.259	.268	.280

first trial; if learning seemingly involves the dropping out of unnecessary responses, the first factor is based on the correlations of the *n*th trial with all of the others. In both cases factors are extracted from successive tables of residuals in the serial order of the learning trials, either forward or backward in time. Similar reasoning holds for developmental data.

An Arbitrary Example

The applicability of diagonal factors to changes over time can be illustrated by an arbitrary example. Table 1 presents a 4×4 matrix of intercorrelations of the hypothetical variable X, which is measured with perfect reliability. Also appearing in this table are the diagonal factors. The first occasion is used to define Factor 1, followed in order by the second, third, and fourth occasions defining Factors 2, 3, and 4.

Although the diagonal method can be applied to any matrix of intercorrelations, it has special properties when used for the simplex. For one thing, the ordering of the variables determines the order in which factors must be extracted if one wishes to portray the special characteristics of the simplex matrix. Secondly, whether factoring starts with the first or the *n*th variable, the factor matrix has one triangle of zero loadings and a second triangle of positive nonzero loadings. The factor matrix directly reflects causes of change. Causes operative before the first measurement occasion, extending backward in time over the history of the organism, are aggregated in the first factor, with the major effect on the first occasion but with effects on all subsequent occasions as well. As new causes of change enter the picture between measurement occasions, the effects occur only on subsequent occasions. New causes appear between each pair of occasions.

The intercorrelations over multiple occasions of each of the two or more measures being studied can be factored by the same methodology, but it is not essential to the accurate description of the intercorrelations that the first factor be defined by the first occasion for each measure. Several different diagonal factorings of the intercorrelations presented in Table 1 are shown in Table 2 for the hypothetical variable Y, which is a variable possibly influenced by X. Although an infinite number of possible rotations are possible mathematically, the four illustrated in Tables 1 and 2 are the appropriate ones for present purposes.

Note that the factoring for variable X could also be extended backward in time. Occasions -1, -2, and -3 would produce columns of factor loadings parallel to those of Factors 3, 2, and 1, respectively, in the Lag 3 factoring of variable Y. Diagonal factoring of X and Y can go on indefinitely both forward and backward in time. For purposes of describing the intercorrelations of each independently of the other, the choice of the occasion on which one decides to start factoring is highly arbitrary.

For purposes of describing a cross-lagged difference, however, choice of occasion is not arbitrary. Factor 1 in X, extracted at Time 1, is identified with a factor in Y, depending on the lag hypothesized, at Time 2, 3, or 4. Depending on this decision, the factor matrix in Table 2 that represents the appropriate lag is selected to represent variable Y. This matrix is then modified by deleting one, two, or three columns from the left-hand side, depending of course on the time lag hypothesized, and adding the same number of columns of zeros on the right. The factor matrix of X is now postmultiplied by the transpose of the

Table 2
Three Alternative Factorings of the Intercorrelations of Table 1

	Factors for Lag 1				Factors for Lag 2				Factors for Lag 3			
Occasion	1	2	3	4	1	2	3	4	1	2	3	4
1	.280	.960	.000	.000	.280	.268	.922	.000	.280	.268	.259	.885
2	.000	1.000	.000	.000	.000	.280	.960	.000	.000	.280	.268	.922
3	.000	.960	.280	.000	.000	.000	1.000	.000	.000	.000	.280	.960
4	.000	.922	.268	.280	.000	.000	.960	.280	.000	.000	.000	1.000

modified Y factor matrix to produce the cross-correlations between the two variables. The discard of factors in Y that preceded the factor identified with Factor 1 in X prior to the matrix multiplication is justified on the basis that they were determined by factors in X that preceded the present Factor 1.

This model can readily be related to the common interpretations of cross-lagged differences. These interpretations have been arranged in order from the one that is closest to the observations and therefore basic to all of the others to the one that clearly requires additional research to establish. The first interpretation, also, serves as the basis for equating an earlier factor in X with a later factor in Y. The interpretations are as follows:

1. Individual differences in X anticipate in time individual differences in Y.
2. Variable X at Time 1 taps the causes for changes in Y that appear subsequent to Time 1.
3. The predominant causal sequence is from X to Y rather than from Y to X.
4. Changes in X cause changes in Y.

A more formal treatment of the argument up to this point follows. Let X be the hypothesized antecedent variable and Y the consequent variable. $\mathbf{F}_x$ is a diagonal factor matrix determined from $\mathbf{R}_x$, and $\mathbf{F}_y$ is a diagonal factor matrix determined from $\mathbf{R}_y$.

$$\mathbf{R}_x = \mathbf{F}_x\mathbf{F}'_x. \tag{1a}$$

$$\mathbf{R}_y = \mathbf{F}_y\mathbf{F}'_y. \tag{1b}$$

Now let $\mathbf{F}_{y_1}, \mathbf{F}_{y_2}, \cdots, \mathbf{F}_{y_n}$ represent factor matrices in which the first column represents the first diagonal factor extracted whether that factor represents Time 1, 2, or n. The $\mathbf{F}_{y_1}$ matrix contains no column of zeros on the right, but all others contain one or more columns of zeros up to $n - 1$ such columns for $\mathbf{F}_{y_n}$.

$\mathbf{R}_{xy} = \mathbf{F}_x\mathbf{F}'_{y_1}$ with time lag of zero between X and Y (no cross-lagged difference observed). (2a)

$\mathbf{R}_{xy} = \mathbf{F}_x\mathbf{F}'_{y_2}$ with time lag of one between X and Y. (2b)

$\mathbf{R}_{xy} = \mathbf{F}_x\mathbf{F}'_{y_n}$ with time lag of $n - 1$ between X and Y. (2c)

Three different sets of possible cross-correlations are shown in Table 3, based upon lags of one, two, and three occasions, respectively. When the lag is one interval, it is seen that the maximum cross-lagged difference occurs at

Table 3
Cross-Correlations Between X *and* Y *for Three Lags*

	Y for Lag 1				*Y* for Lag 2				*Y* for Lag 3			
X	1	2	3	4	1	2	3	4	1	2	3	4
1	.960	1.000	.960	.922	.922	.960	1.000	.960	.885	.922	.960	1.000
2	.922	.960	1.000	.960	.885	.922	.960	1.000	.850	.885	.922	.960
3	.885	.922	.960	1.000	.850	.885	.922	.960	.816	.850	.885	.922
4	.850	.885	.922	.960	.816	.850	.885	.922	.783	.816	.850	.885

that interval; but there is little drop in size at intervals of two or three. When the lag is more than one interval, the maximum cross-lagged difference occurs at the appropriate interval. With zero lag, there are no cross-lagged differences.[1]

Required Additions to the Model

The preceding discussion contains the heart of the model, but it is deficient in three particulars. The matrix multiplication required for Table 3 assumes a one-to-one correspondence between factors in X and Y when the appropriate time lag is selected. This is unrealistic. The most obvious lack is failure to allow for measurement error. When the application is to fallible data, one must substitute reliabilities for the unities in the principal diagonal of the correlation matrix before factoring. Diagonal factoring with reliabilities in the principal diagonal of the R matrix can produce negative residuals that in turn produce negative factor loadings. If the R matrix can be considered to represent a simplex process with little error, residuals from diagonal factoring that would be zero if unities had been used will be very small positive or negative values that can be disregarded.

This means that the matrix multiplications in Equations 1a and 1b still hold, with the exception that the R matrix is defined with reliability estimates rather than with unities in the principal diagonal. This also means that knowledge of reliabilities is an essential requirement in analyzing cross-lagged differences by means of the present model. In this respect the model does not differ from other methods of analysis (see especially, Kenny, 1975).

A second problem is that practically all measures used in the behavioral sciences contain unique nonerror (specific) variance in addition to common factor and error variance. It is logically impossible for the specific variance in X to be related in any way, causally or otherwise, to individual differences in Y. One way or another it is necessary to obtain an estimate of the specificity of both X and Y relative to each other. When circumstances are appropriate, specifics can be estimated by common factor analysis or by multiple regression analysis. At times psychometric analysis suffices. Again, as with measurement error, knowledge of specifics is an essential requirement in analyzing cross-lagged differences. Either reliability differences or specificity differences from one occasion to another can produce completely spurious cross-lagged differences. In Kenny's discussion of a correction for reliability–specificity differences, the two are legitimately merged in the concept of uniqueness.

The presence of specificity does lead to a change in Equations 2a, 2b, and 2c. Two new matrices are required. The first, $\mathbf{H}_x$, is a diagonal matrix consisting of $h_1, h_2, \ldots, h_n$. These values are determined from the proportions of common (h^2) and specific (s^2) variance in the true score variance of X at each time period. The matrix $\mathbf{H}_y$ is defined in parallel fashion. In contrast to $\mathbf{F}_y$, which is modified in accordance with the hypothesized lag, $\mathbf{H}_y$ is independent of lag, that is, $\mathbf{H}_y$ characterizes the measures. If $\mathbf{F}_x$ and $\mathbf{F}_y$ now represent the factors extracted from fallible measures, Equations 2a, 2b, and 2c are rewritten as follows:

$$\mathbf{R}_{xy} = (\mathbf{H}_x\mathbf{F}_x)(\mathbf{F}'_{y_1}\mathbf{H}_y). \quad (3a)$$

$$\mathbf{R}_{xy} = (\mathbf{H}_x\mathbf{F}_x)(\mathbf{F}'_{y_2}\mathbf{H}_y). \quad (3b)$$

$$\mathbf{R}_{xy} = (\mathbf{H}_x\mathbf{F}_x)(\mathbf{F}'_{y_n}\mathbf{H}_y). \quad (3c)$$

The third lack in the general model is only a little less ubiquitous than the presence of measurement error and specificity, but is not as important in its contribution to variance or to model fitting. This source of variance is correlated error that inflates the correlations between variables that are measured on a single occasion. This third component, by definition, can affect only the synchronous correlations among measures and cannot produce spurious cross-lagged differences. Knowl-

[1] These relationships are, of course, the effects of the assumption of a correlation of .96 between true scores on adjacent measurement occasions for both X and Y. This level of correlation was selected as a representative figure for ability measures. Its constant size over both variables and occasions represents the assumption of stationarity that is later discarded. A lower correlation or one that varied from one occasion to another or from one variable to another would have rather different effects.

Table 4
Intercorrelations of Listening and the Composite in Four Grades for 1,430 White Boys and Girls

Test and grade	Listening 5	Listening 7	Listening 9	Listening 11	Composite 5	Composite 7	Composite 9	Composite 11
Listening								
5		.744	.679	.630	.782	.770	.764	.746
7			.751	.683	.760	.830	.820	.802
9				.690	.658	.734	.806	.762
11					.637	.685	.730	.785
Composite								
5						.928	.888	.862
7							.938	.912
9								.930
11								

edge of this source of varance is not, therefore, essential. Absence of this information, however, will prevent one from obtaining a good fit of the synchronous correlations to empirical data.

Aural Comprehension and Intellectual Development

Atkin et al. (1977b) obtained cross-lagged differences that involved a measure of aural comprehension and an intellectual composite of 15 separate cognitive tests. The direction of the differences was that the Listening test predicted the composite more highly than the composite predicted Listening. The differences were also highly significant both statistically and psychologically. Four groups defined by race (black and white) and sex were studied on four occasions, at Grades 5, 7, 9, and 11. Results were highly consistent for the four groups and for the six pairs of occasions, with only one small exception out of 24 comparisons. The size of the difference between the cross-lagged correlations for the four groups in most of the comparisons was between .10 and .20.

In this research the composite that was compared with the Listening test was formed from the following tests: the two tests of School and College Aptitude, the five tests other than Listening from the Sequential Tests of Educational Progress, and the eight narrow information tests that cover heterogeneous areas from the Test of General Information.[2] Similar comparisons for each of the 15 tests in this particular composite with a rotating composite formed in each case from the remaining members of the original set of 16 were also made. Optimum weights obtained from multiple regression and canonical analysis were used to form each composite, but with large *N*s there is little capitalization on chance in the obtained values. No other measure in the set of 16 showed differences that even approached in size or consistency those shown by the Listening test.

To apply the model to these data we decided to combine all correlations that involved white males and females. These two groups contained the larger number of cases (668 males and 762 females). It is also known that the common factor patterns of the 16 tests at each grade level are highly similar for the two white groups (Atkin et al., 1977a). Blacks are not known to differ, but there is more sampling error "noise" in their data. The mean correlations, based on equal weights for the two sexes, are presented in Table 4. When one compares this table with the cross-correlations in Table 3, the asymmetry characteristic of a lag between X and Y is readily apparent.

At present, fitting of the model to the data

[2] These tests are published by the Educational Testing Service (ETS). The data analyzed by Atkin et al. (1977b) were made available by Thomas Hilton and ETS. We owe them thanks for their gracious cooperation.

Table 5
Reliabilities and Specifics for Listening and the Composite in Grades 5, 7, 9, and 11

Statistic	Listening				Composite			
	5	7	9	11	5	7	9	11
Reliability	.743	.817	.753	.693	.937	.981	.957	.947
Specific variance	.106	.108	.084	.066	.000	.000	.000	.000

proceeds one step at a time. The first step requires estimates of reliabilities. In the absence of an experimental design that would make possible test–retest or parallel-forms estimates, it is necessary to assume that the intercorrelations of true scores over the four time periods form simplex matrices. The model developed by Jöreskog (1970) can then be used to estimate reliabilities. To obtain four separate reliabilities from six correlations, it is also necessary to assume that the betas that represent the regression of the true score at Time t on the true score at Time $t - 1$ remain constant. This coefficient was estimated to be .955 for Listening and .968 for the intellectual composite. The estimated reliabilities appear in Table 5. (Also shown are the specific variances that are required later.) With these reliabilities the fit of the simplex to the observed intercorrelations of Listening and the composite, respectively, is excellent. The largest residual for the former is .008 and for the latter .001, and $\chi^2(1) = .80$ and .14, respectively. The reliabilities needed for the diagonal factoring appear to be highly dependable.

Since Listening has the possible causal role in this analysis, the diagonal factoring for this measure starts with Grade 5 and proceeds through Grade 11. Since the lag time between changes in the rank order of persons on Listening and on the composite is unknown, and in the absence of a theory that could guide the choice, the diagonal factoring for the composite uses both Lag 1 and Lag 2. These factors are shown in Table 6.

Although in one sense a digression, a brief discussion of the nature of diagonal factors and factor loadings when reliabilities are inserted in the principal diagonal of the R matrix may be in order. The first factor extracted represents the true score variance of the variable measured on that particular occasion. The factor loadings represent correlations of the fallible measures with this factor. The second factor extracted represents the residual true score variance of the variable measured on the occasion once removed from the one that defined the first factor. The second factor loadings are correlations between residual fallible scores and the residual true score. The factor matrices are not identical with those that would be obtained from R matrices corrected for attenuation with unities in the diagonal. Factor loadings of fallible measures are used to estimate the fallible cross-correlations between Listening and the intellectual composite.

Before proceeding with the matrix multiplication it is necessary to estimate the non-

Table 6
Diagonal Factors in Listening and the Composite

Grade	Listening				Composite							
	1	2	3	4	1	2	3	4	1	2	3	4
5	.862	.000	.000	.000	.243	.937	.000	.000	.243	.231	.908	.000
7	.863	.269	.000	.000	.000	.990	.000	.000	.000	.247	.959	.000
9	.788	.264	.249	.000	.000	.947	.245	.000	.000	.000	.978	.000
11	.731	.228	.217	.245	.000	.921	.237	.207	.000	.000	.951	.207

Table 7
Predicted Cross-Correlations Between Listening and the Intellectual Composite in Grades 5, 7, 9, and 11

	Composite											
	2-year lag				4-year lag				3-year lag			
Listening	5	7	9	11	5	7	9	11	5	7	9	11
5	.764	.806	.771	.750	.740	.782	.797	.775	.752	.794	.784	.763
7	.765	.807	.834	.812	.741	.782	.798	.828	.753	.795	.830	.808
9	.706	.746	.776	.804	.685	.723	.738	.769	.696	.735	.770	.747
11	.662	.700	.723	.747	.642	.678	.691	.718	.653	.689	.719	.742

Note. For the 2-year lag, $\Sigma d = -.211$ and $\Sigma d^2 = .006793$; for the 4-year lag, $\Sigma d = -.041$ and $\Sigma d^2 = .006089$; for the 3-year lag, $\Sigma d = -.103$ and $\Sigma d^2 = .003513$.

error specifics in the Listening test and in the composite. Specific variances for each of the 16 tests were estimated by subtracting the squared multiple correlation between an individual test and the other 15 from the estimated reliability of the test. This was, of course, done at each grade level. The four values of s^2 for Listening, which are presented in Table 5, determined their obverses, the four entries in the H matrix, by means of the relationship $h = (1 - s^2)^{\frac{1}{2}}$. Specific variances of the composite at the four grade levels were set at zero on the basis of the size of the specifics of the individual tests in the composite relative to their reliabilities. As communalities and reliabilities are increased to those for a test 15 times as long as any one component, the communality of the composite approaches very closely its reliability, which in turn approaches unity. Because the intercorrelations are high, specifics tend to be much smaler than measurement error. The result is that specific variance in the composite is trivial in size. It is seen in Table 5 that with an increase in grade and age there is a decrease in the estimated specificity of Listening.

The results of these multiplications are presented in Table 7 for Lag 1 (2 years), Lag 2 (4 years), and an approximation[3] to what a 3-year lag would have produced if the data had been available. In the note underneath the table are descriptive statistics of goodness of fit, namely, the algebraic sum of deviations and the sum of squared deviations between obtained and estimated values for all of the cross-lagged correlations. The discrepancies between obtained and predicted synchronous correlations in the diagonal are ignored in the summations because the model requires a correction for the correlated measurement error that inflates these correlations. These components cannot be estimated independently, at least for the present.

In the absence of a statistical test of goodness of fit one can only conclude that none of the three lags provides a really poor fit and that the approximation to a 3-year lag is clearly the most accurate of the three in terms of the size of the squared deviations. The 4-year lag, on the other hand, is the most accurate in terms of the size of the constant error. The negative sign of this quantity indicates that the model overestimated the observed correlations. If greater weight is placed on the squared deviations as a measure of goodness of fit, and, as a result, the 3-year lag is selected as the most likely one, there are two possible explanations for the constant error. Specific variances may have been slightly underestimated in Listening or in the intellectual composite. Values for the

[3] The equivalent of the first factor that might have been extracted at a 3-year lag if the tests had been administered was obtained by taking the square root of the mean squared factor loadings on the initial factor extracted for the 2- and 4-year lags. From these all other factor loadings can be determined. The resultant is a factor matrix of five columns from which the first two are dropped and a column of zeros added for purposes of the final matrix multiplication.

latter, it will be remembered, were set equal to zero. There is, however, a clear-cut choice between a zero lag and one of from 2 to 4 years. For the former, $\Sigma d = -.24036$ and $\Sigma d^2 = .015351$. In all probability, then, the lag between changes in individual differences in Listening and in the intellectual composite is between 2 and 4 years during the developmental period represented in these data.

We have no theory to support a possible lag of 3 years between individual differences in aural comprehension and in an intellectual composite that both requires reading and includes widely assorted information. The reading disability literature, briefly reviewed by Atkin et al. (1977b), is perhaps relevant. Also, if one accepts extrapolation as a low-level form of theorizing, it seems reasonable to assume that the 4- or 5-year lag between aural and visual comprehension of language between the ages of 2 and 6–7 becomes smaller with increasing age and education. It is surprising, as a matter of fact, to find that the lag remains as great as it seems to be in these data, which cover the period between age 11 and age 17.

It is more difficult to be specific with respect to possible causal inferences that might be drawn from the cross-lagged difference and the model that describes it. It is possible that long-term, continuing experimental manipuation of attentive behavior would affect scores on the Listening test with zero lag and have an effect on scores on the intellectual composite about 3 years later. It is also possible, however, that the causes of change in Listening are beyond any experimental manipulation in any one generation and that these more basic causes affect test scores that require visual comprehension about 3 years after the effect on Listening. It does seem safe to conclude that Listening taps the causes of intellectual development about 3 years earlier than the usual printed intellectual tests between Grade 5 and Grade 11 whether changes in Listening are directly causal or merely anticipate other effects.

The discrepancies in the diagonal would be adequately described by correlated error factor loadings of about .20. These loadings do not seem large relative to the dozen or so possible sources that can be named offhand. These sources are especially potent when testing is done in groups widely scattered geographically with several different test administrators and at somewhat different dates. The test administrators, the examinees, and the measurement settings represent classes of possible sources. Unfortunately, no independent estimate of their impact is available. For the present their assessment is completely circular; that is, they are assessed through discrepancies between estimates and observations.

Inferences From the Model

The simplex process model was introduced by means of an arbitrary example in which stationarity was observed for the two variables and for the four occasions by fixing the correlation between all true scores for adjacent occasions at .96. When the model was applied to data that involved the Listening test and an intellectual composite, the estimate of this correlation for Listening was .955 and for the composite was .968, which represented only a small departure from stationarity. Also, to obtain unique estimates of the reliabilities of the measures on each occasion, it was necessary to assume stationarity of true score regressions from occasion to occasion; that is, stationarity was forced on us by the limitations of the data.

The model is, however, more general than the above discussion indicates. Five occasions would allow estimation of two regressions of one occasion on another. Extrapolation would be required for the other two, but this would be less arbitrary than equating all at the same level. Also, the near equality of the true score regressions for the two variables was not required. The model allows for differences. For example, if correlations of .96 and .94 for X and Y, respectively, had been assumed for the original illustration, a zero time lag would not have produced a symmetrical matrix of cross-correlations. The necessary intercorrelations, factors, and cross-correlations are shown in Table 8: R_y is on the left, F_{y1} is in the center, and $F_x F'_{y1}$ is on the right. Although there are numerical differences between the cross-lagged correlations, these are spurious, being produced by

Table 8
Effects of Reducing the True Score Correlation Between Occasions of Variable Y

Occasion	Intercorrelation 1	2	3	4	Diagonal factor 1	2	3	4	Cross-correlation (zero lag) 1	2	3	4
1	1.000	.940	.884	.831	1.000	.000	.000	.000	1.000	.940	.884	.831
2	.940	1.000	.940	.884	.940	.341	.000	.000	.960	.998	.938	.882
3	.884	.940	1.000	.940	.884	.320	.341	.000	.922	.958	.996	.937
4	.831	.888	.940	1.000	.831	.302	.320	.341	.885	.920	.957	.995

the lack of stationarity from X to Y. The diagonal factoring describes accurately the amount of change from one occasion to another for each variable independently of the other and therefore takes into account the departures from stationarity.

Although stationarity is not required by the model, one must be able to estimate reliabilities and specificities. To fit both intercorrelations and cross-correlations, separate estimates are required. If one is interested only in the cross-correlations, only uniqueness estimates are necessary. With four occasions, reliabilities are estimated from intercorrelations with only one degree of freedom. With less than four occasions reliabilities must be estimated independently of the intercorrelations. Since a simplex process is presumably involved, correlations between any two occasions when corrected for attenuation must be less than unity. Estimation of specificity is a more difficult problem. If there are only two variables, estimates will likely be quite inaccurate. The two synchronous correlations furnish the only information concerning specificity and its obverse, communality.

Stationarity from occasion to occasion may be a reasonable assumption for a great many data, but it is probably unreasonable during periods of rapid development of the organism. Over a long enough time span off-diagonal correlations will probably show a pattern of increasing size. Stationarity from one measure to another is likely to be less tenable, since growth even in related functions can be at different rates. Kenny's (1975) discussion of the problem, as viewed from the present model, misplaces the emphasis. Knowledge of reliability and specificity is primary; stationarity is required for a particular method of correcting for changes in uniqueness. The correction for uniqueness he describes is entirely adequate if the simplex process is stationary for both variables and occasions. Lacking stationarity, however, residual differences will be in part spurious.

The present model also provides insight into the selection of variables to be used in a cross-lagged analysis. The measures must not only be reliable but they must have something in common. One is entitled to be uneasy about the cross-lagged comparisons in Eron, Huesmann, Lefkowitz, and Walder (1972), even though the two cross-correlations differ significantly from each other. In their data the synchronous correlations are too close to zero for any feeling of comfort, and the larger of the two cross-lagged correlations is still quite small. Their measures appear to be highly unreliable, highly specific, or both. There is, however, a mitigating circumstance. Only two time periods were represented, and these were 10 years apart. One additional intermediate occasion would be required as a minimum to resolve these doubts. Although one can say with confidence that no correlation in a simplex matrix should ever be zero, over extended periods of time some of the correlations among measures of moderately stable traits will closely approach zero.

References

Atkin, R., et al. Ability factor differentiation, grades 5 through 11. *Applied Psychological Measurement*, 1977, *1*, 65–76. (a)

Atkin, R., et al. Cross-lagged panel analysis of sixteen cognitive measures at four grade levels. *Child Development*, 1977, *48*, 944–952. (b)

Campbell, D. T., & Stanley, J. C. *Experimental and*

quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Eron, L. D., Huesmann, L. R., Lefkowitz, M. M., & Walder, L. O. Does television violence cause aggression? *American Psychologist,* 1972, *27,* 253–263.

Guttman, L. A generalized simplex for factor analysis. *Psychometrika,* 1955, *20,* 173–192.

Humphreys, L. G. Investigations of the simplex. *Psychometrika,* 1960, *25,* 313–323.

Jöreskog, K. G. Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology,* 1970, *23,* 121–145.

Kenny, D. A. Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin,* 1975, *82,* 887–903.

Rozelle, R. M., & Campbell, D. T. More plausible rival hypotheses in the cross-lagged panel correlation technique. *Psychological Bulletin,* 1969, *71,* 74–80.

Received November 25, 1977 ■

Notice on Author Alterations

Effective January 1979, the procedure for billing authors for alterations will change. Authors of articles in APA journals are customarily charged for alterations on proofs that result from (a) authors' errors or omissions in the manuscript and (b) changes that result from authors' failure to review the edited manuscript or the proof. Until now, alteration charges have been determined by the number of lines affected by changes requested and have been based on 1974 printing rates. Beginning with the proofs for the first issues of 1979, authors' alterations will be determined not only by the number of lines affected but also by the number of figures and pages affected and will be based on current printers' rates.

This change in alteration billing is being made so that alteration charges will more closely reflect the actual cost of changes requested and may mean larger alteration bills for some authors. Authors are reminded that alteration charges apply only when changes are made of proofs and that *there is no charge* for changes made at the manuscript stage.

Psychological Bulletin
1979, Vol. 86, No. 2, 335–337

The Continuing Misinterpretation of the Standard Error of Measurement

Frank J. Dudek
University of Nebraska–Lincoln

Monographs, texts, and guides designed to inform readers about the meanings and interpretations of test scores frequently misinform instead, because the standard error of measurement is misapplied. The standard error of measurement, $\sigma_1(1-r_{1I})^{\frac{1}{2}}$, is an estimate of the variability (i.e., the standard deviation) expected for observed scores when the true score is held constant. To set confidence intervals for true scores given an observed score, the appropriate standard error is that for true scores when observed scores are held constant and estimated by $\sigma_1[r_{1I}(1-r_{1I})]^{\frac{1}{2}}$; and the interval is around the estimated true score rather than around the observed score. Except in the case of perfect reliability, the estimated true score is not the observed score, but is a value regressed toward the mean.

The standard error of estimate, termed the *standard error of measurement* (i.e., $\sigma_{1\infty}$, where the subscript 1 indexes an observed test score and ∞ indexes a presumed *true score* and the order of the subscripts implies that X_1 is predicted from X_∞), seems frequently to be misapplied because many sources tend to promote the erroneous notion that an interval $X_1 \pm \sigma_{1\infty}$ includes the true scores of approximately two thirds of those obtaining scores of X_1 (Educational Testing Service, 1977, p. 16; Lemke & Wiersma, 1976, p. 79; McLaughlin, 1964, p. 18).

A common model (Guilford, 1936, p. 413) for representing the variables involved when observed scores, true scores, and reliability are of concern can be represented by $X_1 = X_\infty + e$ and $X_I = X_\infty + E$, where X_1 and X_I indicate observed scores (e.g., on alternate forms of a test), X_∞ indicates a true score value underlying the observed scores, and e and E represent errors of measurement. It is assumed that (a) true and error components are independent so that $r_{e\infty} = r_{E\infty} = r_{eE} = 0$, (b) the expected value of error components is zero, and (c) $\sigma_e^2 = \sigma_E^2$. As σ_e^2 and σ_E^2 are measures of error variance, the square root of this variance is the standard error of measurement, $\sigma_{1\infty}$, defined above. Under this model it follows that $\sigma_1^2 = \sigma_I^2 = \sigma_\infty^2 + \sigma_{1\infty}^2$; that is, the observed variance of a set of test scores is made up of true variance and error variance. Reliability is given by $r_{1I} = \sigma_\infty^2/\sigma_1^2$ and indicates the proportion of observed score variance that is true variance.

The point that needs to be made about observed scores, true scores, and reliability is that one must distinguish between three different standard errors of estimate, each associated with one of three prediction situations that might be considered. Using deviation scores (i.e., $x = X - M_X$) for simplicity, using $\hat{x}$ to indicate a predicted value, and noting that $r_{1I} = r_{1\infty}^2$, the three standard errors of estimate and the associated prediction to which each applies are

$$\sigma_{1\infty} = \sigma_1(1 - r_{1\infty}^2)^{\frac{1}{2}} = \sigma_1(1 - r_{1I})^{\frac{1}{2}}: \quad \hat{x}_1 = x_\infty. \quad (1)$$

$$\sigma_{\infty 1} = \sigma_\infty(1 - r_{1\infty}^2)^{\frac{1}{2}} = \sigma_1[r_{1I}(1 - r_{1I})]^{\frac{1}{2}}: \quad \hat{x}_\infty = r_{1I}x_1. \quad (2)$$

$$\sigma_{1I} = \sigma_1(1 - r_{1I}^2)^{\frac{1}{2}}: \quad \hat{x}_1 = r_{1I}x_I. \quad (3)$$

The interpretations of all standard errors

Requests for reprints should be sent to Frank J. Dudek, Department of Psychology, 229 Burnett Hall, University of Nebraska, Lincoln, Nebraska 68588.

are analogous; thus, assuming homoscedasticity, (a) $\sigma_{1\infty}$ is the standard deviation of observed scores if the true score is held constant, (b) $\sigma_{\infty 1}$ is the standard deviation of true scores if the observed score is held constant, and (c) σ_{1I} is the standard deviation of X_1 scores if X_I scores are held constant. (It can also be noted that $\sigma_{1I}^2 = \sigma_{1\infty}^2 + \sigma_{\infty 1}^2$.)

The standard error of measurement is commonly reported along with estimates of r_{1I}, and it is important to do so because $\sigma_{1\infty}^2$ is a measure of the amount of error variance that obtains in a set of observed scores. Furthermore, $\sigma_{1\infty}$ as an indicator of measurement error tends to stay constant across populations, whereas r_{1I} varies in magnitude depending on the heterogeneity (i.e., range of talent) represented in the group. Both the reliability coefficient and the standard error of measurement provide useful, descriptive information.

But the interpretation that an interval that extends one standard error of measurement above and one standard error of measurement below an obtained score will include the true scores of approximately two thirds of the individuals who received this obtained score (as implied in the references cited earlier) is in error. The reason is that the prediction implied here is that of a true score *given* an obtained score, and as noted by the prediction in Formula 2, the predicted value of the true score is not the observed score itself, but is an estimate regressed toward the mean. The appropriate standard error of estimate in this situation would be $\sigma_{\infty 1}$, not $\sigma_{1\infty}$.

To illustrate for a set of values in which $\mu_1 = \mu_I = 500$, $\sigma_1 = \sigma_I = 100$, and $r_{1I} = .90$, one finds $\sigma_{1\infty} = 31.623$, $\sigma_{\infty 1} = 30.00$, and $\sigma_{1I} = 43.589$. For all persons who scored 700 on this test, we would infer that the average (i.e., predicted) true score for these persons is 680 and that about two thirds of their true scores lie in the interval 680 ± 30, or between 650 and 710. (On a retest of these individuals one would expect to find two thirds of their retest scores in the interval 680 ± 43.6.)

Standard reference texts (e.g., Guilford, 1936, 1954; Lord & Novick, 1968; Nunnally, 1978) either make or imply the distinctions, of course; but their cautions, caveats, and admonishments are often unheeded, ignored, or misinterpreted in the applied literature.

Guilford (1936) included a footnote:

> Too often one finds the interpretation of a $\sigma_{1\infty}$ misstated. For a given score of 50 when $\sigma_{1\infty}$ is 4, one is likely to read the interpretation that "the probability is two-thirds that the true score lies between 46 and 54." The latter statement implies the prediction of X_∞ from X_1. (p. 414)

On the next page he suggested, "It is correct practice to speak of $\sigma_{1\infty}$ as the *standard error of a raw score* and of $\sigma_{\infty 1}$ as the *standard error of a true score*" (p. 415).

Lord and Novick (1968, pp. 67–68) provided formulations equivalent to our three formulae for the three standard errors of measurement and suggested naming the various standard errors as (a) the standard error of measurement, (b) the standard error of estimation, and (c) the standard error of prediction.

Nunnally (1978), in discussing the standard error of measurement, emphasized that

> One can use it to set confidence zones for obtained scores, but in so doing one must understand that such confidence zones *are not symmetrical about the obtained score*. Thus, although it usually is done in practice, it is incorrect to set the 95 percent confidence zone as equaling two standard errors of measurement below and two above the obtained score. (p. 218)

The first reference to scores in this quotation should have been to true scores rather than to obtained scores. Using the standard error of measurement in the situation in which true score intervals are inferred from obtained scores will not lead to serious error (providing, of course, that regression is taken into account), inasmuch as the standard error in Formula 2 will be less than the standard error in Formula 1 although their values are reasonably close to one another when reliability is high, as seen in the illustration above. Using the standard error in Formula 1, then, even though the standard error in Formula 2 is appropriate will lead to a somewhat liberal interval. But if one desires to set confidence intervals for obtained scores (say on a retest), then the appropriate standard error is σ_{1I}, and using the standard error of measurement in such a situation could lead to a serious underestimation of the interval.

In summary, the standard error of measure-

ment is an estimate of the variability (i.e., standard deviation) of observed scores given a true score and is clearly inappropriate for the situation in which one sets confidence limits for true scores given a fallible, obtained score. For the latter situation one requires the standard deviation of true scores when the observed score is held constant. This standard error is indicated by Formula 2. Equally important is to recognize that the estimated true score, given an observed score, is a value regressed toward the mean, and any confidence interval for true scores will be symmetrical around this regressed value for the true score, not around the observed value.

References

Educational Testing Service. *1977–78 Guide to the Use of the Graduate Record Examinations.* Princeton, N.J.: Author, 1977.

Guilford, J. P. *Psychometric methods.* New York: McGraw–Hill, 1936.

Guilford, J. P. *Psychometric methods.* New York: McGraw–Hill, 1954.

Lemke, E., & Wiersma, W. *Principles of psychological measurement.* Chicago: Rand McNally, 1976.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass: Addison–Wesley, 1968.

McLaughlin, R. F. *Interpretation of test results.* Washington, D.C.: U.S. Government Printing Office, 1964.

Nunnally, J. C. *Psychometric theory.* New York: McGraw–Hill, 1978.

Received December 2, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 338-348

Detecting Cyclicity in Social Interaction

John M. Gottman
University of Illinois at Urbana-Champaign

This article reviews spectral and cross-spectral analytic methods for detecting cyclicity, cross-cyclicity, and lead–lag relationships in continuous data derived from the observation of dyadic interaction. It is found that lead–lag relationships can be assessed using the phase spectrum. Spectral analytic methods are then generalized to categorical observational data, and it is shown that by these methods one can derive the classical information theory definition of social communication and its distribution statistics.

Researchers who study social behavior are discovering that there are occasions when cyclical patterns characterize dyadic interaction, and thus they are searching for statistical techniques that can detect these cycles. The spectral analysis of time-series records was briefly suggested by Luce (1970) as a useful technique for the study of biological rhythms such as heart rate, respiration, REM sleep, and other cyclic biochemical and physiological processes. However, spectral analysis is not widely known to behavioral scientists, and it has yet to be used in the study of social interaction. A recent exception is the work of Hayes and Cobb (Note 1), who observed couples living in a laboratory setting, analyzed cycles of talk and silence using spectral analysis of time-series records, and related an observed cycle to human circadian rhythms.

Researchers who study dyadic social interaction are also interested in the bivariate case in which two time-series records are obtained, one from each of the two interacting organisms; the research question often involves the search for cycles in cross-correlations between the two series. For example, Kendon (1967) reported that when two people converse, the cycles of gaze and gaze aversion interlace, much as do sine and cosine waves. People are out of phase in eye-to-eye contact as a function of who is speaking; in particular, when a person begins speaking he or she looks away from the listener and begins increasing eye-to-eye contact time toward the end of the speech, which acts as an implicit signal for the listener to begin looking away and speaking.

Another example of cross-cyclicity is the work of Brazelton and his associates (e.g., Brazelton, Koslowski, & Main, 1974). Tronick, Als, and Brazelton (1977) studied mother–infant interaction and reported that the infants looked away following periods of maximum involvement with the mother and after a rest period became engaged again. Tronick et al. calculated synchrony and dissynchrony as running correlations between scaled scores of involvement, from maximum positive involvement to maximum negative involvement, but did not employ cross-spectral time-series methods. Cross-spectral analysis would have been an appropriate technique for studying both synchrony and lead–lag relationships between two time series in the Tronick et al. study.

Cross-spectral analysis may have considerable promise for studying interacting physiological systems within an organism. For example, Porges and his associates (Porges, Bohrer, Keren, Cheung, & Franks, Note 2) are using cross-spectral methods to study the linkage between respiration and heart rate. A function called *coherence* obtained from cross-spectral analysis is the equivalent of the

---

This research was supported by Grant MH 29910 from the U.S. Public Health Service. The author wishes to acknowledge discussions of the phase spectrum with Duane Steidinger and Robert Bohrer.

Requests for reprints should be sent to John M. Gottman, 609 Psychology Building, University of Illinois, Champaign, Illinois 61820.

square of the correlation between the two physiological systems as a function of their relative lag. Porges et al. found that the coherence between respiration and heart rate is related to cognitive attentional processes. Hyperactive children had low coherence between respiration and heart rate; low doses of methylphenidate had positive influence on cognitive performance and social behavior, whereas higher doses often resulted in lethargy. Porges and his associates are testing the model that deficits in linkage between the respiratory system and the cardiovascular systems are related to the attentional problems of hyperactive children and that low doses of methylphenidate mediate to increase the coherence between systems, thereby affecting cognitive functioning.

Because time-series techniques are not widely known to psychologists, this article reviews the spectral and cross-spectral analysis of continuous data. The present research also derives the new result that the slope of the phase spectrum of any two stationary processes can be used to detect lead–lag relationships. Lead–lag relationships are useful in making inferences about which series is, in some sense, driving the other. One application of lead–lag relationships is a redefinition of the concept of dominance in social interaction as an asymmetry in predictability in the time domain (Gottman, in press). This definition of dominance using cross-spectral analysis subsumes a range of observations about dominance across species. For example, the beta male in a group of monkeys is more responsive to the behavior of the alpha male than conversely (Maslow, 1936); that is, the behavior of the beta male is more predictable from past behavior of the alpha male than conversely.

Most researchers of social interaction collect categorical rather than continuous observational data (e.g., Hutt & Hutt, 1970; Lewis & Rosenblum, 1974). There are currently no statistical techniques for detecting cycles in one sequence and cycles between two sequences for categorical data over time. Categorical data collected over time can always be transformed to continuous time-series data; for example, for every block of k time units the local probability of each category can be computed, which produces a continuous variable for each category. For a discussion of categorical data types in observational research, see Gottman and Bakeman (in press).

In this article, I derive extensions of spectral time-series methods to categorical data. One result of these extensions is the derivation of the commonly used information theory definition of communication, summarized by Wilson (1975) as follows:

> Communication has been defined as the process by which behavior of one individual alters the probability of behavioral acts in other individuals . . . . In words, the conditional probability that act X_2 will be performed by individual B given that A performed X_1 is not equal to the probability that B will perform X_2 in the absence of X_1. (p. 194)

This is an important definition for the study of sequences in social interaction because it suggests the notion that a behavior in one organism has social communicative value to the extent that it reduces uncertainty in predicting the behavior of another organism. This definition is now widely used to detect sequences in dyadic interaction (for reviews, see Gottman & Bakeman, in press; Gottman & Notarius, 1978; Sackett, 1977).

Another result of the extension of spectral time-series methods to categorical data in this article is the demonstration of the validity (and limitations) of a statistical test of significance between conditional and unconditional probabilities recently suggested by Sackett (1977). After the information theory definition of communication is derived, spectral and cross-spectral methods are used to suggest how lead–lag relationships and cycles can be detected in categorical time series.

The Continuous Case

Granger and Hatanaka (1964) noted that the first time series subjected to spectral analysis were those that had a cycle with one dominant frequency, such as the 11-year oscillation in sunspot data and the annual cycle in meteorological data. They wrote,

> It was felt that if one could determine the amplitude period and phase of a sine curve sufficiently accurately and subtract this from the data, then the remainder ought to be an independent, random series. When, in fact, this was done and the remainder was still found to be somewhat too smooth, it was natural to re-use the current predominant idea of the cause of the

smoothness and to look for yet further sine curves to fit to the data. (pp. 4–5)

The model for a time series, X_t, was therefore a weighted sum of sine and cosine curves with an uncorrelated random remainder; if the number of observations, $n = 2q + 1$, is odd, one can write

$$X_t = A_0 + \sum_{i=1}^{q} (A_i \cos 2\pi f_i t + B_i \sin 2\pi f_i t) + \epsilon_t, \quad (1)$$

where $f_i = i/n$ is the ith harmonic of the fundamental frequency $1/n$. Fourier analysis makes it possible to derive least squares estimates for the coefficients:

$$\hat{A}_0 = \bar{X} = \frac{1}{n} \sum X_t;$$

$$\hat{A}_i = \frac{2}{n} \sum_{t=1}^{n} X_t \cos 2\pi f_i t;$$

$$\hat{B}_i = \frac{2}{n} \sum_{t=1}^{n} X_t \sin 2\pi f_i t.$$

This decomposition of a time series into component frequencies met with some initial success. For example, Whittaker and Robinson (1924) showed that the brightness of a variable star could be decomposed into two component frequencies, and they thus determined that the variable star was a binary star.

It would be useful to have some function that peaked at frequency bands that made major contributions to the variance of the series. For an infinite number of observations, the variance of the series at each frequency, f_i, is called the spectral density function, f. For a sample of n points it is called the *periodgram:* $I(f_i) = (1/8\pi)(A_i^2 + B_i^2)$. Because the sine and cosine terms in Equation 1 form an orthogonal set of functions, it can be shown that the variance of the time series is partitioned into independent parts by the periodogram:

$$\frac{1}{n} \sum (X_t - \bar{X})^2 = \frac{1}{2n} \sum I(f_i).$$

Early work on the spectral analysis of time series suggested that the periodogram was precisely the function that would peak at frequencies that contributed major portions to the variance of the time series; in fact, Schuster (1898) suggested that the periodogram be calculated and that its peaks be used to detect cycles. Subsequently, problems with spurious peaks led to the construction of significance tests for the periodogram (for a review of these tests, see Jenkins & Priestley, 1957). However, these tests were not adequate because the periodogram has some very poor statistical properties.

If the sample autocovariance at lag k is defined as

$$C_k = \frac{1}{n} \sum_{t=1}^{n-k} X_t X_{t-k},$$

then C_k is an unbiased estimator of the population autocovariance (Hannan, 1967), and it can be shown (Box & Jenkins, 1970, p. 45) that the periodogram is given by

$$I(f) = \frac{1}{2\pi}(C_0 + 2 \sum_{k=1}^{n-1} C_k \cos 2\pi f k),$$

where $0 \leq f \leq \frac{1}{2}$, which expresses that the periodogram is the Fourier transform of the sample autocovariance function. This implies that the periodogram is also easily calculated from the sample autocovariances; thus at first the problems of spectral time-series analysis appeared to be solved.

Unfortunately, although the periodogram does converge to the spectral density function, f, it does not converge uniformly; that is, its variance around f does not decrease to zero as n, the number of observations, increases (Hannan, 1967, pp. 52–53). In fact, Bartlett (1948) showed that the limit of the variance of the periodogram as n increases is $\sigma^4 f^2$, where σ^2 is the variance of the series. The failure of the periodogram led Tukey (1967) to make the following reflection:

> If we dealt with problems involving the superposition of a few simple periodic phenomena, as do astronomers interested in binary stars and related problems, we can learn much from the periodogram. Sadly, however, almost no one else has this kind of data. As a result the periodogram has been one of the most misleading devices I know. (p. 25)

A dramatic illustration of Tukey's point is the periodogram of a series of random numbers, called white noise. White noise, like white light, is composed of all frequencies with equal intensities, and therefore its periodogram should be a straight line. Jenkins and Watts (1968) showed that the periodogram of white

noise is not only not a straight line but continues to oscillate wildly as the number of observations is increased. However, the spurious peaks of the periodograms of each sample of white noise occur in random places on the frequency domain, and this provides the key to solving the problems of the periodogram. The average of many periodograms obtained from many samples of the same white noise process in fact tends toward a straight line as the number of observations in each sample increases.

This observation led Bartlett (1948) to suggest that the time series can be segmented and that a periodogram can be averaged across all segments. Bartlett showed that the averaged periodogram would coverage uniformly to the spectral density. Jenkins (1967) demonstrated that Bartlett's suggestion is equivalent to estimates of the form

$$f(f_i) = \frac{1}{2\pi}\left[C_0 + 2\sum_{j=1}^{n-1} \lambda_j(f_i)C_j \cos 2\pi f_i j\right]. \quad (2)$$

The function of $\lambda(f_i)$ is called a *spectral window*, and it weights the autocovariance function to ensure uniform convergence. Parzen's (1967) result is important because to implement Bartlett's suggestion would require an extremely long time series, whereas Jenkins's suggestion can be implemented with shorter time series, assuming that the window weighting function is suitably chosen. The most commonly used spectral window is the Tukey–Hanning window (Blackman & Tukey, 1958): $\lambda_j = 1 + \cos(\pi j/m)$, where m is an arbitrary integer, usually chosen so that $m < n/3$. (See Parzen, 1967, for a discussion of various spectral windows.) Thus a weighted Fourier transform of the autocovariance function does converge uniformly to the spectral density. Jenkins and Watts (1968) showed that the distribution of the intensity estimates at each frequency of the periodogram "will be very nearly a χ_2^2 regardless of the distribution of the [time-series] process" (p. 233). For the Tukey–Hanning window, the equivalent degrees of freedom must be modified (Granger & Hatanaka, 1964, pp. 59–64; Jenkins & Watts, 1968, pp. 248–257). In this article the term *spectrum* refers to the weighted periodogram.

An illustration of the spectrum may clarify its relationship as the Fourier transform (with an appropriate spectral window) of the autocovariance function. If the time series is a second-order autoregressive process,

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t, \quad (3)$$

where ϵ_t is an uncorrelated, random series and $\phi_1^2 + 4\phi_2 < 0$, then the behavior of the series will appear periodic. The constraints on ϕ_1 and ϕ_2 occur because periodicity only occurs when the roots of the characteristic equation of the process are imaginary (Box & Jenkins, 1970, p. 59). Note that this time series will not be deterministically periodic, as is a sine wave; there is a random component to the periodicity. In this case the autocovariance function will be a single-frequency damped sine wave.[1] Figure 1 is a plot of the autocorrelation function and spectrum of a simulated second-order autoregressive model. The spectrum shows only one peak[2]; a fourth-order autoregressive process is capable of representing a process with two peaks, and so on.

The relationship between the autocorrelation and spectrum of the process represented by Equation 3 is intuitively clear. If the time series is periodic, the autocorrelation should increase at multiples of the period. For

---

[1] The expression for the theoretical autocorrelation function is

$$\rho_k = \frac{[\text{sgn}(\phi_1)]^k d^k \sin(2\pi f_0 k + F)}{\sin F},$$

where sgn $= +1$ if ϕ_1 is positive and sgn $= -1$ if ϕ_1 is negative. The factor d is called the *damping factor*, f_0 is called the *frequency*, and F is called the *phase*. These factors are related to the model parameters as follows:

$$d = [(-\phi_2)^{\frac{1}{2}} \text{sgn}(\phi_1)];$$

$$\cos 2\pi f_0 = \frac{|\phi_1|}{2(-\phi_2)^{\frac{1}{2}}};$$

$$\tan F = \frac{1+d^2}{1-d^2}\tan 2\pi f_0.$$

[2] The spectrum of a second-order autoregressive process can be written in closed form as

$$p(f) = 2\sigma_a^2/[1 + \phi_1^2 + \phi_2^2 - 2\phi_1(1-\phi_2)\cos 2\pi f - 2\phi_2 \cos 4\pi f],$$

where $0 \le f \le \frac{1}{2}$. The spectrum reflects the periodic behavior of the second-order autoregressive process when the roots of its characteristic equation are complex.

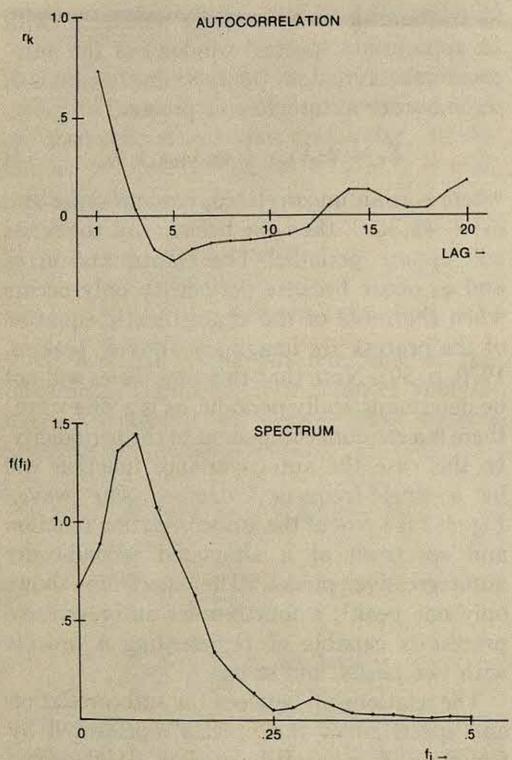

Figure 1. Autocorrelation function (r_k is the autocorrelation at lag k) and spectrum, $f(f_i)$, of one realization of a second-order pseudoperiodic time series. ($X_t = 1.1X_{t-1} - .5X_{t-2} + e_t$.)

example, for monthly wholesale wheat prices, the correlation between months 12 months apart (June with June, July with July, etc.) should be higher than that between months in different seasons. This relationship should fall off across years, and so the autocorrelation should resemble the damped sine wave in Figure 1. Since there is only one 12-month cycle, one would expect the spectrum (the weighted Fourier transform of the autocovariance function) to show only one peak. If the series had two cycles, the autocorrelation function would appear similar in shape (but more complex), and the spectrum would have two peaks.

Note that one cannot reconstruct the original time series simply by knowledge of the spectrum. This is true because very different series can be produced simply by adjusting the relative phases of the component frequencies; the phase of a sine wave determines its amplitude at time zero. Phase is a particularly important concept in the bivariate case.

For two time series, the generalization is not difficult. In fact, the Fourier transform (with suitable window) of the cross-covariance between the two time series is the cross-spectrum. The cross-spectrum has several components, a phase spectrum, and a cross-amplitude spectrum. The phase spectrum indicates

> whether the frequency components of one series lead or lag the same frequency components in the other series, and the cross-amplitude spectrum shows whether the amplitude of the component at a particular frequency in one series is associated with a large or small amplitude at the same frequency in the other series. (Jenkins & Watts, 1968, pp. 342–343)

The coherence is a function similar to the square of a correlation coefficient and is defined as the ratio of the square of the cross-spectrum divided by the product of the spectra of the individual series[3]; for two series, X_t and Y_t,

$$K_{xy}(f_i) = \frac{|f_{xy}(f_i)|^2}{f_{xx}(f_i)f_{yy}(f_i)}. \tag{4}$$

Distribution properties of these functions are discussed in Jenkins and Watts (1968, chap. 9); the properties for these functions with the Tukey–Hanning window are discussed in Granger and Hatanaka (1964, chap. 5). A coherence of one means that prediction is perfect from one series to another for all frequencies; a coherence of zero means that it is impossible to predict one series from the other. The prediction is of amplitude covariations in the two series, with no indication of lead–lag relationships, so that a complete description of relationships requires the phase spectrum as well as the coherence. If the coherence has one major peak, then the bulk of the correlation between the two processes is confined to a particular frequency band. If it is essential to predict correlations at major frequency bands of series Y_t, the coherence can be investigated at frequencies that have peaks in the spectrum of Y_t. An

---

[3] An alternative approach for specifying the relationship between two time series in the time domain, as opposed to in the frequency domain, is called *transfer function analysis* and is discussed by Box and Jenkins (1970).

alternative, suggested by Porges et al. (Note 2), is to compute one statistic called the *weighted coherence*, which is an estimate of the amount of variation in one series that can be accounted for by variation in the other:

$$\sum_i k_{xy}(f_i) f_{xx}(f_i) / \sum_i f_{xx}(f_i).$$

They wrote,

> Conceptually the coherence may be thought of as a time-series analogue of the omega-squared . . . . or the amount of variance accounted for by the influence of one series on the other. Therefore, the coherence times the spectral density estimate of heart rate activity at each frequency . . . . would describe the amount of heart rate activity which could be accounted for by respiration, i.e., the shared variance of heart rate and respiration. (p. 5)

If the cross-covariance is $C_{xy}(t)$, the unweighted cross-spectrum is the Fourier transform of the cross-covariance:

$$f_{xy}(f) = \frac{1}{2\pi} \sum_{-\infty}^{\infty} e^{i2\pi f t} C_{xy}(t), \quad \text{where} \quad i = (-1)^{\frac{1}{2}}.$$

This complex number can be written as a real part plus an imaginary part: $f_{xy}(f) = C + iQ$. The phase spectrum is defined as

$$\phi_{xy}(f) = \arctan \frac{Q}{C}; \tag{5}$$

C is called the cospectrum and Q the quadrature spectrum, and they measure the covariance between in-phase and out-of-phase components, respectively.

The slope of the phase spectrum determines the time-lag and the lead–lag relationships between the two series. For example, if one time series, $X(t) = \epsilon(t)$, is white noise with variance σ^2 and the other series is $Y(t) = X(t + L)$, then L is the *lead time* and Y leads X by L time units later. Since $X(t)$ is white noise, the covariance of $X(t)$ and $Y(t)$ is

$$\begin{aligned} C_{xy}(t) &= E[X(s)Y(s + t)] \\ &= E[X(s)X(s + t + L)] \\ &= \begin{cases} \sigma^2 & \text{at} \quad t = -L \\ 0 & \text{otherwise} \end{cases} . \end{aligned}$$

The cross-spectrum is the Fourier transform of the cross-covariance. Assuming $\sigma = 1$, this

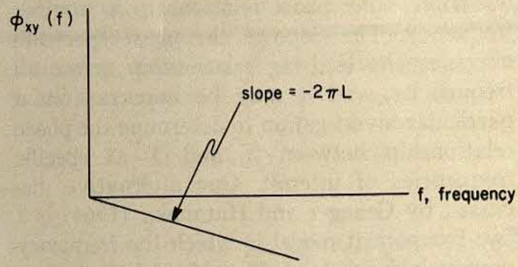

Figure 2. Phase spectrum, $\Phi_{xy}(f)$, when $Y(t)$ leads $X(t)$ by a constant time, L.

gives

$$\begin{aligned} f_{xy}(f) &= \frac{1}{2\pi} \sum_{-\infty}^{\infty} e^{i2\pi f t} C_{xy}(t) \\ &= \frac{\sigma^2}{2\pi} e^{-i2\pi f L} = \frac{\sigma^2}{2\pi} (\cos 2\pi f L - i \sin 2\pi f L), \end{aligned}$$

where $i = (-1)^{\frac{1}{2}}$. The phase spectrum is given by

$$\begin{aligned} \phi_{xy}(f) &= \arctan \left(\frac{Q}{C} \right) = \arctan \left(\frac{-\sin 2\pi f L}{\cos 2\pi f L} \right) \\ &= \arctan(-\tan 2\pi f L) = -2\pi f L. \end{aligned}$$

Therefore the phase spectrum will be a straight line that passes through the origin with negative slope proportional to the time lag, L (see Figure 2).

More generally, lead–lag relationships can be estimated by testing the significance of the slope of the least squares linear regression approximation to the phase spectrum.[4] It is important to note that this method does not give complete information; X and Y may be periodic at a particular frequency and have a constant phase relationship at that frequency

---

[4] The phase spectrum shown in Figure 2 can be shown to hold for any two stationary processes that differ by a constant time lag. If $Y(t) = X(t + L)$, then the Fourier transform of $Y(t)$ is

$$F[Y(t)] = \int_{-\pi}^{\pi} Y(t) e^{iwt} dt = \int_{-\pi}^{\pi} e^{iwt} X(t + L) dt.$$

If one lets $u = t + L$, then

$$F[Y(t)] = \int_{-\pi}^{\pi} e^{iw(u-L)} X(u) du = e^{-iwL} \int_{-\pi}^{\pi} e^{iwu} X(u) du;$$

$$F[Y(t)] = e^{-iwL} F[X(t)].$$

Hence the Fourier transform of $Y(t)$ is the Fourier transform of $X(t)$ multiplied by the phase shift $e^{-i\phi}$, where $\phi = -wL$.

but some other phase relationship at another frequency. The slope of the phase spectrum averages the lead–lag relationship across all frequencies, and it may be important in a particular investigation to determine the phase relationship between X and Y at specific frequencies of interest. One alternative discussed by Granger and Hatanaka (1964) is a two-component model in which the frequency domain is divided in half and lead–lag relationships are assessed separately for slow and rapid components. For these calculations, computer programs are available in most universities that have the University of California, Los Angeles biomedical series (Dixon, 1974, pp. 517–582, Programs 2T, 3T, and 4T).

The Categorical Case

In the categorical case two series, X_t and Y_t, are set equal to one if the characteristics that they represent are observed and equal to zero otherwise. The unbiased estimator of the cross-covariance, lagged k units in time is

$$C_{xy}(k) = \frac{1}{n-k}\sum_{1}^{n-k}(X_t - \bar{X})(Y_{t+k} - \bar{Y}).$$

For categorical data, $\bar{X}$ and $\bar{Y}$ are the unconditional probabilities, p_x and p_y, that X_t and Y_t are one in $n - k$ observations, so that

$$\begin{aligned} C_{xy}(k) &= \frac{1}{n-k}\sum_{1}^{n-k}(X_t - p_x)(Y_{t+k} - p_y) \\ &= \frac{1}{n-k}\Big[\sum_{1}^{n-k} X_t Y_{t+k} - p_y \cdot \sum_{1}^{n-k} X_t \\ &\quad - p_x \cdot \sum_{1}^{n-k} Y_{t+k} + p_x p_y (n-k)\Big] \\ &= \frac{1}{n-k}\Big[\sum_{1}^{n-k} X_t Y_{t+k} - p_y p_x (n-k) \\ &\quad - p_x p_y (n-k) + p_x p_y (n-k)\Big]; \end{aligned}$$

$$C_{xy}(k) = \frac{1}{n-k}\Big[\sum_{1}^{n-k} X_t Y_{t+k} - p_y p_x (n-k)\Big]. \quad (6)$$

The sum in Equation 6 is simply the number of lagged-k (1, 1) pairs. Note that by definition, the conditional probability that Y is equal to one given that X was equal to one k time units ago, $p_k(Y|X)$, is simply the number of (1, 1) pairs at lag k divided by the number of occurrences of $X = 1$ in $n - k$ observations. If one denotes the number of (1, 1) pairs at lag k as $M_{xy}(k)$, then, from the definition of the lagged conditional probability, it follows that $p_k(Y|X) = M_{xy}(k)/p_x(n-k)$. Therefore, the number of (1, 1) pairs at lag k is

$$M_{xy}(k) = \sum_{1}^{n-k} X_t Y_{t+k} = p_k(Y|X)(p_x)(n-k).$$

Substituting this back into Equation 6 gives

$$C_{xy}(k) = p_x[p_k(Y|X) - p_y], \quad (7)$$

as the categorical equivalent of the cross-covariance.

This function is proportional to the information theory definition of communication assessed as the difference between conditional and unconditional probabilities.

To derive the distribution of the covariance, the variance of the covariance can be computed as follows:

$$\begin{aligned} C_{xy}(k) &= p_x[p_k(Y|X) - p_y]; \\ C_{xy}(k) - \bar{C} &= p_x p_k(Y|X) - p_x \overline{p_k(Y|X)} \\ &= p_x[p_k(Y|X) - \overline{p_k(Y|X)}]; \\ \operatorname{var}[C_{xy}(k)] &= p_x^2\{\operatorname{var}[p_k(Y|X)]\}. \end{aligned}$$

Under the null hypothesis of no relationship between the two categorical time series, X_t and Y_t, $p_k(Y|X) = p_y$, and the variance of the unconditional probability of a dichotomous variable that is not autocorrelated is $p_y(1 - p_y)/m$ (Siegel, 1956, p. 40), where m = the number of observations used to calculate p_y. For the covariance $C_{xy}(k)$, $m = n - k$, and the result is

$$\operatorname{var}[C_{xy}(k)] = p_x^2 p_y(1 - p_y)/(n-k).$$

Since under the null hypothesis, $C_{xy}(k)/SD[C_{xy}(k)]$ is normally distributed with mean zero and unit variance ($N[0, 1]$) (Box & Jenkins, 1970), one has

$$\frac{C_{xy}(k)}{SD[C_{xy}(k)]} = \frac{p_x[p_k(y|x) - p_y]}{[p_x^2 p_y(1-p_y)/(n-k)]^{\frac{1}{2}}} \sim N(0, 1);$$

$$Z = \frac{p_k(y|x) - p_y}{[p_y(1-p_y)/(n-k)]^{\frac{1}{2}}} \sim N(0, 1). \quad (8)$$

This is a derivation of a statistic that was recently proposed by Sackett (1977).

An estimate of the error introduced in Equation 8 by autocorrelation in each series, under the null hypothesis of no cross-correlation, can be obtained by using the expression for the variance of the cross-correlation under the null hypothesis given by Box and Jenkins (1970, p. 377):

$$\operatorname{var}[r_{xy}(k)] \simeq \frac{1}{n-k}[1 + \sum_{j=1}^{\infty} r_{xx}(j) r_{yy}(j)]$$

$$\simeq \frac{1}{n-k}(1+\delta).$$

Thus, the variance of the cross-correlation would be $1/(n-k)$ if there were no autocorrelation. To estimate the quantity δ, rewrite the autocorrelations using

$$r_{xx}(k) = C_{xx}(k)/C_{xx}(0):$$

$$\delta = \sum_{1}^{\infty} r_{xx}(j) r_{yy}(j)$$

$$= \frac{1}{C_{xx}(0) C_{yy}(0)} \sum_{1}^{\infty} C_{xx}(j) C_{yy}(j).$$

Now substitute the quantity for the covariance from Equation 7:

$$\delta = \frac{p_x p_y}{p_x(1-p_x) p_y(1-p_y)} \sum_{1}^{\infty} [p_j(x|x) - p_x] \times [p_j(y|y) - p_y].$$

If one assumes that the quantity in the sum decreases exponentially with increasing lag and one denotes

$$\theta = [p_1(x|x) - p_x][p_1(y|y) - p_y],$$

then

$$\delta = \frac{1}{(1-p_x)(1-p_y)} \cdot \frac{\theta}{(1-\theta)}.$$

Delta is a maximum when the conditionals are one and a minimum when the conditionals equal the unconditionals:

$$\delta_{\max} = \frac{1}{1-(1-p_x)(1-p_y)}; \quad \delta_{\min} = 0.$$

The cross-spectral density function for categorical data can be written as the Fourier transform of the cross-covariance (weighted by a suitable window), and this function will behave in a fashion similar to the continuous case. The generalizations are obtained by applying Equation 2 to Equation 7: The cross-spectrum is

$$f_{xy}(f_i) = \frac{1}{2\pi}[C_{xy}(0)\lambda_0(f_i) + 2 \sum_{j=1}^{n-1} \lambda_j(f_i) C_{xy}(j) \cos 2\pi f_i j].$$

The spectrum of X_t is

$$f_{xx}(f_i) = \frac{1}{2\pi}[p_x(1-p_x)\lambda_0(f_i) + 2 \sum_{j=1}^{n-1} \lambda_j(f_i) C_{xx}(j) \cos 2\pi f_i j].$$

The spectrum of Y_t is

$$f_{yy}(f_i) = \frac{1}{2\pi}[p_y(1-p_y)\lambda_0(f_i) + 2 \sum_{j=1}^{n-1} \lambda_j(f_i) C_{yy}(j) \cos 2\pi f_i j].$$

The lambdas are the Tukey–Hanning weights (Blackman & Tukey, 1958).

To summarize, Equation 7 is the categorical equivalent of the cross-correlation, and if $X = Y$, of the autocorrelation. If cyclicity exists in a series of categorical data with one major cycle, then $C_{xy}(k)$ should behave as a damped sine wave of Figure 1, and the spectrum should show one peak. An examination of the spectrum, which is the weighted Fourier transform of Equation 7, reveals major cycles in the categorical series. The coherence and phase spectrum are similarly generalized, and the slope of the phase spectrum detects lead–lag relationships that span all component frequencies. Computationally, all these statistics can be calculated simply by inputting each series as a binary zero–one time series.

To illustrate the relationship between continuous and dichotomous spectral time-series statistics, one example that compares statistics for continuous data and the same data dichotomized around the mean is presented.

Example

The data in this example are derived from coding a videotape of a married couple working

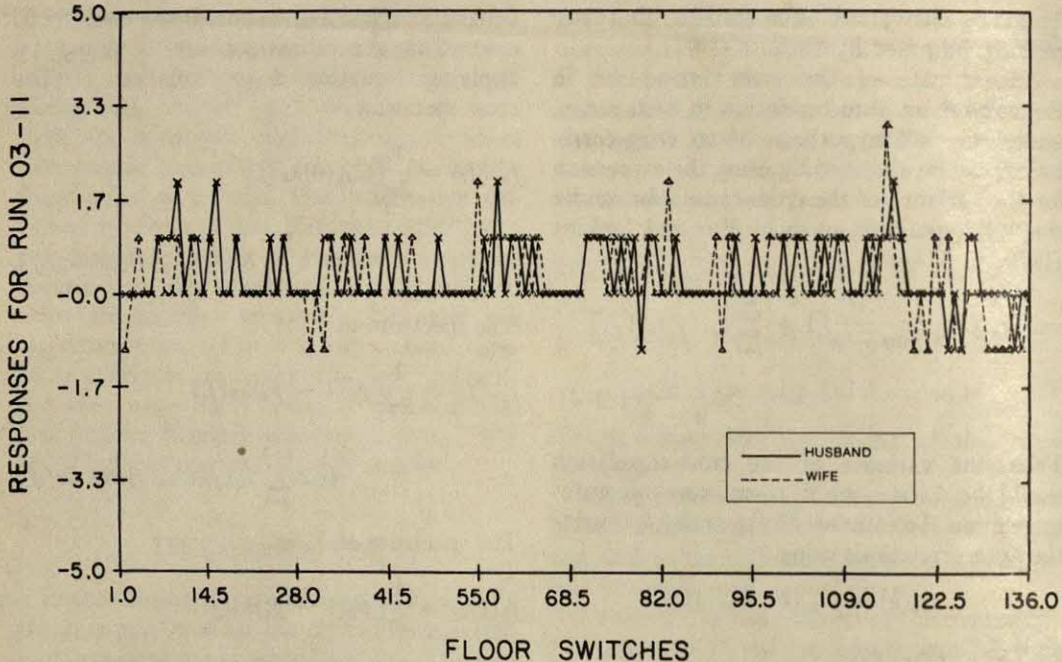

Figure 3. Positivity of behaviors of one couple on an improvised conflict task.

on an improvised conflict task. The coding system and the method for generating the time series from categorical data are described in Gottman, Markman, and Notarius (1977). The graphs displayed in Figure 3 represent a tally of positive minus negative nonverbal behavior coded from voice tone, facial expressions, and body cues. The unit plotted on the abscissa is the "floor switch," that is, the set of utterances before one person gives up the floor to the other.

These data were transformed to categorical data by dichotomizing around the mean of each series, and phase spectra and the coherences were calculated for both the discrete and the continuous cases using Tukey–Hanning

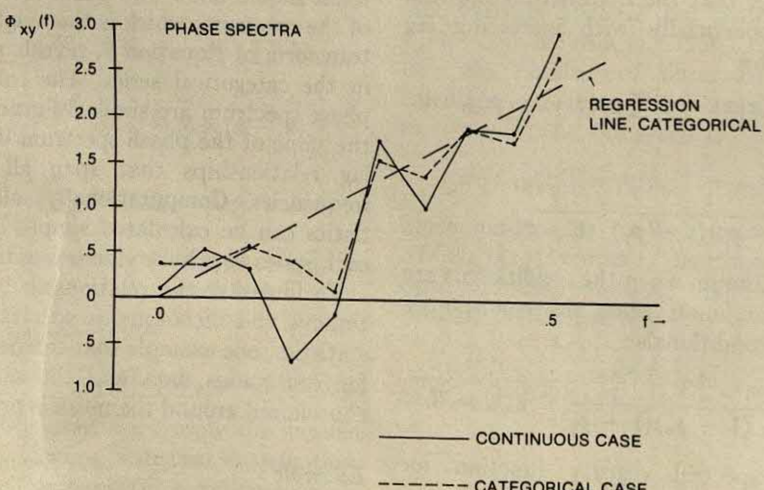

Figure 4. Phase spectra for continuous and dichotomous case of couple in Figure 3; $\Phi_{xy}(f)$ = phase spectrum; f = frequency.

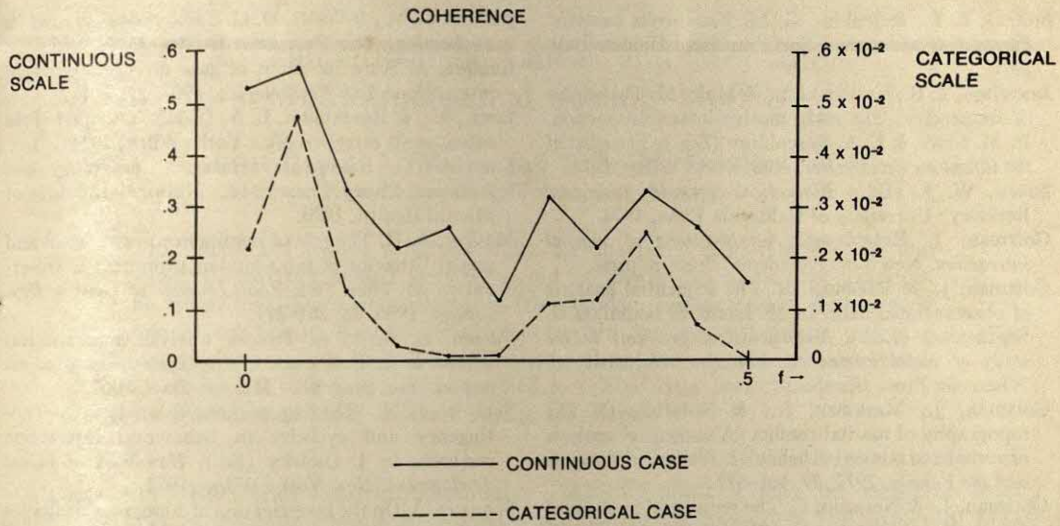

Figure 5. Coherence spectra for continuous and dichotomous cases of couple in Figure 3; f = frequency.

weights and the Fast-Fourier transform program available at the University of Illinois (SOUPAC programs). Figures 4 (see Equation 5) and 5 (see Equation 4) present a comparison of these two statistics for the continuous and categorical cases. The phase spectra are nearly identical and have very similar regression lines that in both cases are interpreted as the wife leading the husband, with a constant lag equal to the slope of the regression line. The slope is .31 for the continuous case and .25 for the categorical case.

The coherence for the categorical case is much lower, which is not surprising because so much information about strength of association is lost by dichotomizing. However, the important aspect of the coherence is the location of peaks, and one can see that the coherence for the categorical case has a shape similar to that of the continuous case. The two highest peaks (at $f = .1$ and $f = .4$) are the same for both cases, so that information about cyclicity in the strength of association across series is preserved.

Conclusion

Spectral and cross-spectral time-series methods were reviewed in this article for continuous data, and interpretations were discussed for the spectrum, the coherence, the weighted coherence, and the phase spectrum. These methods were also extended to categorical data. The extension made it possible to derive the information theory statistic for comparing conditional with lagged unconditional probabilities and for exploring the limits of the z-score test as a function of autocorrelation. Subsequent investigations should generate stochastic time-series data by using known autoregressive-moving average models with seasonal components (Box & Jenkins, 1970) and by comparing continuous and dichotomous analyses. The methods proposed in this article need to be applied to a range of problems, and their ability to describe patterns in data across time and to fail to detect patterns in known random data needs to be assessed empirically.

Reference Notes

1. Hayes, D. P., & Cobb, L. *The temporal organization of long-term social interaction.* Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1976.
2. Porges, S. W., Bohrer, R. E., Keren, G., Cheung, M. N., & Franks, G. J. *Respiratory sinus arrythmia: A time-series model assessing the influence of methylphenidate on vagal tonus.* Unpublished manuscript, University of Illinois, 1977.

References

Bartlett, M. S. Smoothing periodograms from time-series with continuous spectra. *Nature*, 1948, *161*, 686–687.

Blackman, R. B., & Tukey, J. W. *The measurement of power spectra.* New York: Dover, 1958.

Box, G. E. P., & Jenkins, G. M. *Time-series analysis: Forecasting and control.* San Francisco: Holden–Day, 1970.

Brazelton, T. B., Koslowski, B., & Main, M. The origins of reciprocity: The early mother-infant interaction. In M. Lewis & L. A. Rosenblum (Eds.), *The effect of the infant on its caregiver.* New York: Wiley, 1974.

Dixon, W. J. (Ed.). *Biomedical computer programs.* Berkeley: University of California Press, 1974.

Gottman, J. *Experimental investigations of marital interaction.* New York: Academic Press, in press.

Gottman, J., & Bakeman, R. The sequential analysis of observational data. In M. Lamb, S. Soumi, & G. Septhenson (Eds.), *Methodological problems in the study of social interaction.* Madison: University of Wisconsin Press, in press.

Gottman, J., Markman, H., & Notarius, C. The topography of marital conflict: A sequential analysis of verbal and nonverbal behavior. *Journal of Marriage and the Family*, 1977, *39*, 461–477.

Gottman, J., & Notarius, C. The sequential analysis of observational data using Markov chains. In T. Kratochwill (Ed.), *Strategies to evaluate change in single subject research.* New York: Academic Press, 1978.

Granger, C. W. J., & Hatanaka, M. *Spectral analysis of economic time series.* Princeton, N.J.: Princeton University Press, 1964.

Hannan, E. J. *Time-series analysis.* London: Methuen, 1967.

Hutt, S. J., & Hutt, C. *Direct observation and measurement of behavior.* Springfield, Ill.: Charles C Thomas, 1970.

Jenkins, G. M. General considerations in the analysis of spectra. In E. Parzen (Ed.), *Time-series analysis papers.* San Francisco: Holden–Day, 1967.

Jenkins, G. M., & Priestley, M. B. The spectral analysis of time-series. *Journal of the Royal Statistical Society* (Series B), 1957, *19*, 1–12.

Jenkins, G. M., & Watts, D. G. *Spectral analysis and its applications.* San Francisco: Holden–Day, 1968.

Kendon, A. Some functions of gaze direction in social interaction. *Acta Psychologica*, 1967, *26*, 1–47.

Lewis, M., & Rosenblum, L. A. (Eds.). *The effect of the infant on its caregiver.* New York: Wiley, 1974.

Luce, G. G. *Biological rhythms in psychiatry and medicine.* Chevy Chase, Md.: National Institute of Mental Health, 1970.

Maslow, A. H. The role of dominance in the social and sexual behavior of infra-human primates: I. Observation at Vilas Park Zoo. *Journal of Genetic Psychology*, 1936, *48*, 261–277.

Parzen, E. Notes on Fourier analysis and spectral windows. In E. Parzen (Ed.), *Time-series analysis papers.* San Francisco: Holden–Day, 1967.

Sackett, G. P. The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In J. Osofsky (Ed.), *Handbook of infant development.* New York: Wiley, 1977.

Schuster, A. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terra Magnum*, 1898, *3*, 13–41.

Siegel, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

Tronick, E. D., Als, H., & Brazelton, T. B. Mutuality in mother-infant interaction. *Journal of Communication*, 1977, *27*, 74–79.

Tukey, J. W. An introduction to the calculations of numerical spectrum analysis. In B. Harris (Ed.), *Spectral analysis of time series.* New York: Wiley, 1967.

Whittaker, E. T., & Robinson, G. *The calculus of observations.* London: Methuen, 1924.

Wilson, E. D. *Sociobiology: The new synthesis.* Cambridge, Mass.: Belknap Press, 1975.

Received December 5, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 349–354

A Comparison of Linear and Monotone Multidimensional Scaling Models

David G. Weeks and Peter M. Bentler
University of California, Los Angeles

Multidimensional scaling solutions under the linear and monotone (metric and nonmetric) distance models were compared for several monotone and nonmonotone distortions. Data were generated from random configurations, over a wide range of conditions. Results indicate that, when its assumptions are met, the linear model performs best. When linearity assumptions are not met, the monotone model and the linear model applied to ranked data perform equally well. Recommendations based on these results are offered.

Models for Euclidean multidimensional scaling (MDS) of a single symmetric matrix can be summarized as $h_{ij} = f\{[\Sigma_k(a_{ik} - a_{jk})^2]^{\frac{1}{2}}\}$, where h_{ij} is the observed (dis)similarity between stimuli i and j, and a_{ik} is the projection of stimulus i on dimension k. Although in principle f may be virtually any real valued function, in practice there are two main types. In the linear MDS model, $h_{ij} = bd_{ij} + c$, and the parameter b is an arbitrary scale factor that may be ignored. This linear model is commonly referred to as metric MDS. The monotone model specifies

$$h_{ij} \overset{m}{=} d_{ij},$$

that is, the distances are required only to stand in the same monotone (m) rank order as the dissimilarities. More explicitly, the model requires $d_{ij} \leq d_{kl}$ iff $h_{ij} < h_{kl}$. This monotone model is generally referred to as nonmetric MDS.

This research was supported in part by U.S. Public Health Service Grants MH24149 and DA00017. The University of California, Los Angeles, Office of Academic Computing provided over $2,000 in computer resources required for the simulations and analyses.

We wish to thank J. D. Carroll, E. W. Holman, J. B. Kruskal, F. W. Young, and several anonymous reviewers for helpful comments.

Requests for reprints should be sent to David G. Weeks or Peter M. Bentler, Department of Psychology, University of California, Los Angeles, California 90024.

At this point it may be useful to briefly outline certain aspects of the history of MDS, as this relates to the problems investigated in the current article. The first method for MDS was developed by Torgerson (1952, 1958) and was based on a theorem by Young and Householder (1938). This method solves for the linear model under the Euclidean metric. Nonmetric MDS was introduced by Shepard (1962a, 1962b); this allowed solutions for the monotone model. Kruskal (1964a, 1964b) defined an explicit error function for this model that was minimized by the method of steepest descent. Cooper (1972) developed a method for solving the linear model using a method related to that used by Kruskal, and Bentler and Weeks (1978) developed restricted, hypothesis-testing methods for the linear model. Since its introduction in 1962, the monotone model has come to dominate the field.

The continued use of the monotone model would imply that its superiority over the linear model had been established. Although it has been shown that the monotone model works quite well (e.g., Shepard, 1962b), the superiority of the monotone model over the linear model remains undemonstrated. This point can be made clearer by an examination of a classic study in the field. Ekman (1954) obtained similarity ratings of 14 spectral colors. These ratings were linearly transformed to a 0–1 scale. Then, on the assumption that these values were equivalent to

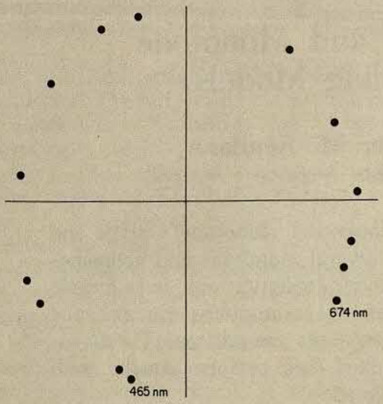

Figure 1. Ekman (1954) color data analyzed by the linear model. (The stimuli were colored lights from 434 nm [violet] to 674 nm [red].)

cosines or correlations between vectors representing the stimuli, the configuration was obtained by a relatively primitive type of factor analysis. The resulting solution was five color clusters in five factors. Shepard (1962b) reanalyzed these data under the monotone MDS model with his new method. He obtained a solution in two dimensions, resembling the color circle. Although Shepard's solution was clearly more appropriate than Ekman's, it was not compared with the linear distance model. Consequently, Shepard's solution does not verify the superiority of the monotone model over the linear model. Indeed, our own reanalysis of the Ekman data under the linear model (see Figure 1) revealed a solution indistinguishable from Shepard's monotone solution.

One would expect the monotone model to be dramatically superior only when the data are substantially nonlinear, but it is possible that most data typically used for MDS are approximately linear on distances. Furthermore, large deviations from "monotonicity" would also tend to be the largest deviations from linearity. "Clearly the success of such an undertaking [mapping a proximity matrix into Euclidean space] depends upon the selection of the proper distance function; that is, the function that will transform the proximity measures into Euclidean distances" (Shepard, 1962a, p. 127). Shepard (1962b) gave several examples of analyses of distances distorted by nonlinear monotone functions. In all cases his method recovered the true configuration, as well as the shape of the distorting function. However, these results were not compared with solutions under the linear model, so that it is not possible to conclude that monotone MDS was markedly superior to linear MDS. It is our purpose in this article to compare results under the linear and monotone models and to explore the robustness of the linear model. We are concerned only with the limited problem of comparing the linear and monotone models in Euclidean exploratory multidimensional scaling; no attention is paid the other important problems (such as how to handle missing data, analysis of non-Euclidean metrics, parameter-constrained scaling methods, transformation of initial solutions to aid interpretation, scaling individual-difference data, the value of alternative nonlinear optimization methods, the relative merit of alternative initial starting configurations, methods for avoiding or evaluating local minima, etc.) that have no relevance to this comparison.

The Problem of Comparison

Attempting to compare metric and nonmetric MDS presents certain serious problems. Results of the comparison should be useful in applied situations, and for that purpose, analysis of real data is advised. In applied situations, however, it may not be possible to determine which solution is "best." In some cases, one solution may make more sense than another, but this is by no means assured (this, of course, presumes that a difference will be found). In many cases, the data might be nearly linear on the true distances and would thus provide no test of the power of nonmetric MDS. Generating distances from a known configuration, and distorting them by a known monotone function, seems more promising. However, the problems of comparison do not end there by any means. Stress, Kruskal's (1964a, 1964b) measure of poorness of fit, is inadequate for the present purpose because it means something different in either case. In the linear model, every deviation from linearity contributes to

stress; in the monotone model, only deviations from monotonicity are counted. The same solution—provided it does not give perfect fit—has to have a higher stress value if stress is measured in terms of linear rather than monotone regression.

It must be acknowledged that certain important determinants of the quality of an MDS solution are not amenable to Monte Carlo, computer simulation techniques. In particular, an acceptable solution is one that makes sense in terms of the particular stimulus domain. Such judgments are usually subjective, often necessarily so. Criteria are also idiosyncratic to the particular data situation. Nonetheless, there is an important component of the quality of a solution that is amenable to quantification and hence to Monte Carlo techniques: the nearness of the estimated parameters to the true parameters of the model. For the sake of simplicity, we consider only the elements of the projection matrix as parameters. The problem reduces to one of finding a measure to characterize the discrepancy between an obtained and a true configuration. The squared correlation between the distances generated by the two configurations is an appropriate measure. In a distance model, differences in orientation and location of the origin are irrelevant, because distances are invariant under such transformations. Usually central dilation is irrelevant, affecting only the scale of the distances. The correlation coefficient is, of course, scale invariant. The correlation of distances is natural in the sense that the error-free component of an observation used in MDS is a distance. The squared correlation has been used before in similar applications (Girard & Cliff, 1976; MacCallum & Cornelius, 1977; Rabinowitz, 1976; Sherman, 1972; Young, 1970). A viable alternative would be a measure of association between two matrices, as developed by Lingoes and Schönemann (1974; Schönemann & Carroll, 1970). We prefer the squared correlation in this case, because it is simpler in both calculation and interpretation (the proportion of variance of one set of distances accounted for by the other). Of course, it must be recognized that since the distances within a set are not independent, standard statistical sampling theory is not relevant to understanding the squared correlation in this context.

A critical problem in MDS is determination of the correct dimensionality (Shepard, 1974). Probably the most common method now in use involves extracting only as many dimensions as can reasonably be interpreted. This method, of course, is not amenable to Monte Carlo studies. A promising objective method involves matching obtained stress curves with those previously obtained by Monte Carlo techniques (e.g., Spence & Graef, 1974; Wagenaar & Padmos, 1971). Unfortunately, these studies employed only the monotone model, and thus their results cannot be used to compare the monotone model with the linear model.

A Simulation Study

Method

Conditions were chosen to cover the range of conditions found in most applications of MDS. Three levels of number of points ($m = 10$, 20, or 30) and four levels of true dimensionality ($t = 1$, 2, 3, or 4) were completely crossed. Elements of the configuration matrices were obtained from a uniform pseudorandom number generator with a range of 1.4, 1.2, 1.0, or .8 for the first through fourth dimensions, respectively. Within each m,t combination there were two independent replications.

Data were obtained by generating true distances from the configuration and by adding random error to the distances. For cases in which this led to negative values, a constant was added such that the smallest value was zero. These values were then transformed by several known functions to produce the sets of data that were analyzed. Error was drawn from a normal distribution with a mean of zero and variances, expressed as proportions of the variance of the true distances, of .25, .75, and 2.0. (The generator used was routine GGNRF [*IMSL Library*, 1975].) Five distorting functions were chosen. The first, $h_{ij} = d_{ij} + e_{ij}$, where h is the data, consisted of no distortion—a linear relationship between data and distances. This was in a sense a control condition by which the severity of the other distortions could be judged. Two distortions, $h_{ij} = (d_{ij} + e_{ij})^4$ and $h_{ij} = (d_{ij} + e_{ij})^{\frac{1}{4}}$, were intended to exemplify severe, monotone distortions. The fourth distortion, $h_{ij} = \text{rank}\ (d_{ij} + e_{ij})$, was particularly important for two reasons: First, the ranked data contain only that information used by the monotone model; second, this condition can always be obtained in real situations by ranking the data, since $\text{rank}\ [f(d + e)] = \text{rank}\ (d + e)$, where f is monotone. The last distortion was $h_{ij} = |d_{ij} + e_{ij} - w|$,

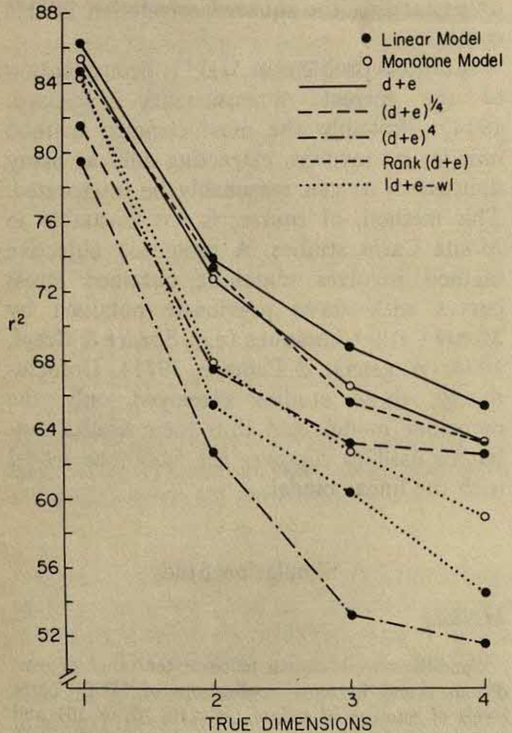

Figure 2. Effect of model, distortions, and number of dimensions on the squared correlation (r^2); $d =$ distance; $e =$ error; $w =$ a constant.

where w is the 20th percentile of $d + e$. This distortion was included more for curiosity than for practical value. We wished to gain some insight into the behavior of the two models when monotonicity was violated.

Thus, 3 (number of points) × 4 (dimensions) × 2 (replications) × 3 (levels of error) × 5 (distortions) = 360 sets of data that were generated. All sets of data were analyzed under the linear model. Under the monotone model, the various monotone distortions have no differential impact. Therefore, only data produced by the linear function and the nonmonotone function were analyzed under the monotone model. (Actually, we performed complete analyses for a small subset of the conditions and confirmed that this was true in practice as well as in theory.) All analyses were done with the KYST program (Kruskal, Young, & Seery, 1973), which uses Kruskal's (1964a, 1964b) method. Starting configurations were obtained by Torgerson's (1958) method. Stress 1, with the primary approach to ties, was specified. Solutions were obtained in from one to six dimensions, but the obtained dimensionality was not allowed to exceed the true dimensionality plus three. This resulted in 10,584 solutions. For each solution, stress and the squared correlation between true and recovered distances were recorded.

Results

Of primary interest were differences in recovery of the true distances between the linear and monotone models. The mean squared correlation for the seven distortion-model conditions, for the cases in which recovered dimensionality equaled true dimensionality, is plotted in Figure 2. (All correlations were positive.) The linear model, with no distortion, was best overall. The rank condition with the linear model and the no-distortion condition with the monotone model were virtually identical and were nearly as good as the first condition. The two other monotone distortions with the linear model were clearly worse than the monotone analysis of ranked data, with which they should properly be compared. For the nonmonotone distortion, the monotone model was superior to the linear model except in the one-dimensional condition.

The effects of error, dimensions, and the number of points on the squared correlation are shown in Figure 3. Again, these data are

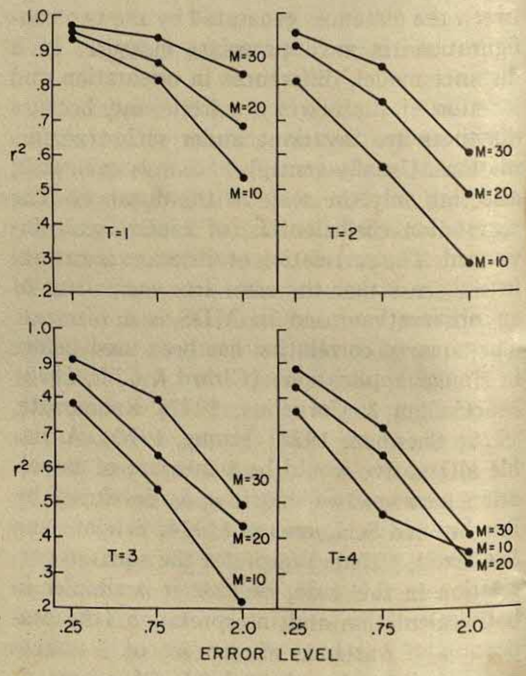

Figure 3. Effect of error, number of points, and number of dimensions on the squared correlation (r^2); T corresponds to true dimensionality; $M =$ number of points.

for recovery in the true dimensionality of the configuration. The squared correlation decreases as error and number of dimensions increase and increases with the number of points. The effects of number of points and number of dimensions reflect the degree of overdetermination of parameters. There were no systematic interactions of error with type of analysis.

The effects of the various factors were examined for the data in which the dimensionality was overestimated or underestimated by one. There were no coherent patterns of results in these data that were different from the data displayed in Figures 2 and 3. Therefore, detailed summaries of these data are omitted.

Discussion

Two aspects of the monotone multidimensional scaling model can be conceptually distinguished. The first is that only the ordinal properties of the data are retained, and the second is that a best fitting monotone function is derived. The results of the present study strongly indicate that only the first aspect has any real practical importance. The ordinal constraints among a matrix of distances are strong enough to determine a highly metric solution (Kruskal & Shepard, 1974; Shepard, 1966, 1974; Young, 1970), and this is true whether the function relating data to recovered distances is linear or monotone. The virtual identity in the squared correlation between the linear, rank conditions and the monotone, no-distortion conditions bears this out. Further, since the squared correlation values in the linear, no-distortion conditions were only slightly higher than either the ranked, linear conditions or the monotone conditions, it seems that the actual interval properties of the data add little over the ordinal constraints.

The other aspect of the monotone model, the function derived from monotone regression, has a serious drawback not shared by the linear model. It is well-known that a problem arises when there are clusters of points such that the within-cluster distances are all smaller than the between-clusters distances. Use of the monotone model can then lead to degenerate solutions, in which the clusters collapse into single points (Shepard, 1974). This can happen because monotonicity is preserved when all small distances are set to zero and all larger distances are set equal to one another. The linear model is not similarly susceptible to degeneracy, because monotone equating of distances creates deviations from linearity, which add to the size of the loss function. Consequently, the rank version of the linear model (rank-linear model) should also be able to avoid such degenerate solutions.

The generally outstanding performance of the rank-linear model is not particularly surprising when one considers that the use of ranks involves using more information from the original data in the rank-linear model than from the original data in the monotone model. In the rank-linear model, the ranks represent a certain distance function for the data; in the monotone model, the distance function is recovered subject only to the weaker constraint that $d_{ij} \leq d_{kl}$ iff $h_{ij} < h_{kl}$. Thus the rank-linear model operates in a data metric analogous to Spearman's rank-order correlation coefficient, whereas the monotone model operates in a data metric analogous to Kendall's tau. In the closely related area of monotone principal-components analysis (e.g., Kruskal & Shepard, 1974), the use of ranked data similarly yields excellent results at a large savings in cost (Woodward & Overall, 1976).

Based on our results, the known problems with monotone regression, and cost considerations, one may wish to approach multidimensional scaling analyses in the following way. First, the scaling is performed under the linear model with no transformation of data. If there are no systematic monotone biases, the solution should be optimal as well as inexpensive. A scatterplot of data to recovered distances must be examined; if systematic nonlinearities are present, the data should be ranked and reanalyzed with the linear model. If no appreciable clustering of points is detected, a monotone scaling would be a suitable (but more expensive) alternative to the second analysis. This approach to

MDS is not appropriate when there is a nonmonotone relation between data and distances. Nonmonotone functions are in general not one to one, they do not necessarily have inverses, and an adequate recovery of the function is not possible. If severe nonmonotonicities exist in the data, it would be neither appropriate nor profitable to employ any distance model. If one were employed, however, the monotone model would seem to be a reasonable choice. It was more robust than the linear model to violations of monotonicity, at least in our data.

In conclusion, the monotone model has the advantage of robustness over the linear model, since, in this study, it performed better than the linear model for all systematic monotone and nonmonotone distortions. The linear model, on the other hand, has the advantage of conceptual simplicity and greater computational efficiency and avoids the danger of degeneracy. The rank-linear model appears to offer the advantages of both. The only computation it requires over the linear model is an initial ranking of data. The ranking eliminates all systematic monotone nonlinearities, whereas the linear analysis avoids the potential of degeneracy due to monotone regression.

References

Bentler, P. M., & Weeks, D. G. Restricted multidimensional scaling models. *Journal of Mathematical Psychology*, 1978, *17*, 138–151.

Cooper, L. G. A new solution to the additive constant problem in metric multidimensional scaling. *Psychometrika*, 1972, *37*, 311–322.

Ekman, G. Dimensions of color vision. *Journal of Psychology*, 1954, *38*, 467–474.

Girard, R. A., & Cliff, N. A Monte Carlo evaluation of interactive multidimensional scaling. *Psychometrika*, 1976, *41*, 43–64.

IMSL Library (Vol. 1, Ed. 5). Houston, Tex.: International Mathematical and Statistical Libraries, 1975.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, *29*, 1–27. (a)

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964, *29*, 115–129. (b)

Kruskal, J. B., & Shepard, R. N. A nonmetric variety of linear factor analysis. *Psychometrika*, 1974, *39*, 123–157.

Kruskal, J. B., Young, F. W., & Seery, J. B. *How to use KYST, a very flexible program to do multidimensional scaling and unfolding*. Murray Hill, N.J.: Bell Laboratories, 1973.

Lingoes, T. C., & Schönemann, P. H. Alternative measures of fit for the Schönemann-Carroll matrix fitting algorithm. *Psychometrika*, 1974, *39*, 423–427.

MacCallum, R. C., & Cornelius, E. T., III. A Monte Carlo investigation of recovery of structure by ALSCAL. *Psychometrika*, 1977, *42*, 401–428.

Rabinowitz, G. A procedure for ordering object pairs consistent with the multidimensional unfolding model. *Psychometrika*, 1976, *41*, 349–373.

Schönemann, P. H., & Carroll, R. M. Fitting one matrix to another under choice of central dilation and a rigid motion. *Psychometrika*, 1970, *35*, 245–255.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function: I. *Psychometrika*, 1962, *27*, 125–140. (a)

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function: II. *Psychometrika*, 1962, *27*, 219–246. (b)

Shepard, R. N. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 1966, *3*, 287–315.

Shepard, R. N. Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 1974, *39*, 373–421.

Sherman, C. R. Nonmetric multidimensional scaling: A Monte Carlo study of the basic parameters. *Psychometrika*, 1972, *37*, 323–335.

Spence, I., & Graef, J. The determination of the underlying dimensionality of an empirically obtained matrix of proximities. *Multivariate Behavioral Research*, 1974, *9*, 331–341.

Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, *17*, 401–419.

Torgerson, W. S. *Theory and method of scaling*. New York: Wiley, 1958.

Wagenaar, W. A., & Padmos, P. Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 101–110.

Young, F. W. Nonmetric multidimensional scaling: Recovery of metric information. *Psychometrika*, 1970, *35*, 455–473.

Young, G., & Householder, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 1938, *3*, 19–22.

Woodward, J. A., & Overall, J. E. Factor analysis of rank-ordered data: An old approach revisited. *Psychological Bulletin*, 1976, *83*, 864–867.

Received December 5, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 355–360

Comment on Olson: Choosing a Test Statistic in Multivariate Analysis of Variance

James Stevens
University of Cincinnati

This article questions Olson's claim that the Pillai-Bartlett statistic (V) is superior to Wilks's λ (W) and the Hotelling-Lawley trace (T) for general use in multivariate analysis of variance because of much greater robustness against unequal covariance matrices. It is shown by a sampling of studies from the literature that the example Olson used to demonstrate superiority had extreme subgroup variance differences, which occur very infrequently. For subgroup variance differences much more likely to occur, it is shown that the actual Type I error rates for V, T, and W are very similar. For concentrated noncentrality structures with covariance heterogeneity, it is recommended that any of these three statistics be used, since the slight robustness advantage V has is offset by the greater power of T and W in these situations. For diffuse structures, V is the clear choice as test statistic.

Olson (1976) argued recently, based on robustness and power considerations, that of four multivariate test statistics available (Roy's largest root [R], the Hotelling-Lawley trace [T], Wilks's λ[W], and the Pillai-Bartlett trace [V]), V should generally be used. Olson's argument was mainly based on his statement that V is much more robust than are the other three to violations of the homogeneity-of-covariance-matrices assumption in multivariate analysis of variance.

To illustrate the much greater robustness of V, Olson gave the following example from his 1976 article: Consider a three-group experiment with five subjects per group and six dependent variables at nominal $\alpha = .05$. Then if one group is sampled from a population with variances 36 times as great as those of the other groups, the actual Type I error rate for $V = .09$; but the error rate is .49 for W, .58 for T, and .62 for R. This result does show V to be dramatically more robust than T, W, or R. However, a representative sampling of Olson's (1973) dissertation results indicates that V is only dramatically more robust for extreme subgroup variance differences, as in the above example. Table 1, which presents the sampling of actual Type I error rates from Olson's dissertation for nominal $\alpha = .10$ and .05, shows that for smaller subgroup variance differences the differences in error rates for V, T, and W are very small. The differences in error rates for T, W, and V for nominal $\alpha = .10$ are less than 2%; for W versus V, they are less than about 1%. The differences in error rates for T, W, and V for the more extreme nominal value of .05 are essentially the same. In general, they are very small ($\leq$ 2.3% for T vs. V and $\leq$ 1.5% for W vs. V).

Eight of the nine cases in Table 1 in which the differences in error rates are larger (marked with superscript b) for both nominal values correspond to very large subgroup variance differences on all variables (36I, that is, the population variances of all variables in one group are 36 times greater than the variances of those variables in the remaining groups). For nominal $\alpha = .10$, the actual error rates are at least twice the nominal value. For nominal $\alpha = .05$, the actual error rates are at least three times the nominal value. Therefore, under 36I the error rates for all statistics are far from the nominal value, although relatively speaking V is much more robust.

Requests for reprints should be sent to James Stevens, Teachers College Building, University of Cincinnati, Cincinnati, Ohio 45221.

Table 1
Actual Type I Error Rates for Four Multivariate Test Statistics Under Heterogeneous Covariance Matrices for Nominal α = .10 and .05

No. variables	No. groups	N	D[a]	Roy's largest root R .10	Roy's largest root R .05	Hotelling trace T .10	Hotelling trace T .05	Wilks's W .10	Wilks's W .05	Pillai-Bartlett trace V .10	Pillai-Bartlett trace V .05
2	3	5	4I	138	89	139	80	143	70	137	65
2	3	5	9I	185	140	180	123	178	109	167	102
2	3	5	36I	252	198	234	176	229	166	218	151
2	3	5	C(4)	122	72	129	65	128	61	124	66
2	3	5	C(9)	139	92	141	82	144	73	139	75
2	3	5	C(36)	156	104	149	98	159	91	157	87
2	3	10	4I	133	87	128	73	124	67	125	63
2	3	10	9I	168	111	153	101	150	100	144	90
2	3	10	36I	209	142	182	130	174	126	168	117
2	3	10	C(4)	114	67	116	65	112	63	111	61
2	3	10	C(9)	123	75	120	70	118	69	119	71
2	3	10	C(36)	126	78	126	76	127	77	123	78
2	3	50	36I	185	129	157	100	154	99	153	96
2	3	50	C(36)	136	81	127	72	127	72	124	71
2	6	5	4I	160	100	138	79	128	71	121	62
2	6	5	9I	243	188	205	142	192	121	164	98[b]
2	6	5	36I	355	298	299	243	268	218	231	170[b]
2	6	5	C(4)	133	77	125	79	119	76	113	75
2	6	5	C(9)	165	110	152	100	142	102	134	98
2	6	10	4I	145	96	122	81	115	78	113	69
2	6	10	9I	209	163	166	135	161	127	158	114
2	6	10	36I	286	238	223	191	207	184	197	162[b]
2	6	10	C(4)	121	79	118	81	113	79	111	73
2	6	10	C(9)	137	94	121	98	122	96	124	94
2	6	10	C(36)	161	117	142	113	139	115	137	109
3	3	10	4I	165	92	140	88	135	86	126	75
3	3	10	9I	217	150	185	130	176	125	166	111
3	3	10	36I	274	202	224	188	217	184	197	153[b]
3	3	10	C(4)	108	56	110	63	111	64	108	63
3	3	10	C(36)	115	74	120	74	121	80	117	78
3	3	20	4I	167	117	146	89	149	86	145	83
3	3	20	9I	221	162	173	128	171	123	165	109
3	3	20	36I	264	203	199	160	195	152	190	143
3	3	20	C(4)	131	89	131	84	132	78	131	70
3	3	20	C(36)	145	100	142	95	142	95	139	90
3	6	10	4I	197	127	146	100	137	92	129	77
3	6	10	36I	434	373	300	261	273	224	233	186[b]
3	6	10	C(4)	143	83	124	79	124	78	119	74
3	6	10	C(36)	171	109	157	107	156	101	147	100
6	3	10	36I	542	462	443	358	369	299	277	187[b]
6	6	50	36I	632	550	270	223	249	206	232	182[b]
10	3	10	36I	830	752	752	688	673	564	336	165[b]
10	3	10	C(36)	122	63	126	67	115	58	116	63
10	3	50	36I	487	434	263	211	234	186	212	152[b]
10	6	10	C(36)	135	84	130	92	129	91	136	73

[a] D = extent of subgroup variance differences: 4I means population variances of all variables in one group are four times greater than the variances of those variables in the remaining groups; C refers to concentrated differences. Thus, C(4) means the population variance on only one variable in one group is four times greater than the variance of that variable in the remaining groups.
[b] Cases in which the differences in error rates for T versus V are greater than 2.5% and for W versus V are greater than 1.5%.

Table 2
A Sampling of Subgroup Variance Differences on a Variety of Variables from Nine Different Studies

Wright, 1975

Stanford Achievement subtest	Traditional Boys	Traditional Girls	Open Boys	Open Girls
Word meaning	1.85	.90	.72	.49
Paragraph meaning	2.82	1.49	1.14	1.06
Spelling	1.82	2.25	1.06	1.42
Language	1.82	1.54	1.17	1.02
Arithmetic comprehension	1.66	.76	.66	.64
Arithmetic concepts	1.56	1.00	1.02	.74
Social studies	1.54	.76	.77	.40
Science	1.66	1.39	.92	.72
Arithmetic applications	4.80	1.49	.83	1.17

French, Brownell, Graziano, & Hartup, 1977[b]

Variable	1	2	3	4
Blocks	102.8	384.9	160.5	222.9
Falls	.79	1.7	3.2	.64
Discrepancy	38.7	23.7	20.2	.70

Gardner & Schumacher, 1977[d]

Variable	1	2	3	4	5	6
1	4.5	8.3	11.0	10.9	14.7	14.7
2	7.0	13.0	5.1	13.3	14.8	12.9
3	8.1	15.8	12.7	17.6	22.3	22.8
4	12.5	7.2	7.6	12.6	10.8	18.0

Stevens, 1972[a]

Vocational personality inventory variable	1	2	3	4
Realistic	9.31	6.95	7.26	5.38
Intellectual	22.41	17.41	18.18	18.33
Social	14.60	14.44	13.96	15.50
Conventional	8.99	7.06	7.47	3.72
Enterprising	7.16	8.02	8.94	8.84
Artistic	16.23	15.36	18.99	16.43
Status	4.00	5.60	5.48	6.48
Aggressive	16.36	15.31	11.68	12.89

Smith, Gnanadesikan, & Hughes, 1962[c]

Variable	1	2	3	4
Pigment creatine	15.05	8.22	19.29	4.19
Chloride	3.67	5.80	3.09	.88
Choline	28.54	16.30	30.56	85.77

Novince, 1977

Self-report variable	Behavioral rehearsal	Control	Behavioral rehearsal and cognitive restructuring
Avoidance	296.89	281.89	261.42
Negative evaluation	499.36	123.25	226.80
Social interaction	120.85	226.82	198.62
Self-report inventory of anxiousness	188.16	381.36	499.87

(*table continued*)

Table 2 (*continued*)

Finn, 1974[e]

Variable	1	2	3	4	5	6	7
1	1.07	2.00	.286	1.00	2.97	.42	.78
2	6.21	2.69	.143	.70	5.44	.57	1.39
3	7.36	4.86	1.244	1.00	6.14	1.06	5.64
4	4.13	8.00	3.620	.70	10.79	2.70	6.22
5	3.36	4.78	2.000	.30	5.66	1.19	2.22
6	.98	1.53	.290	0	3.32	.64	1.36

Calsyn, Spengler, & Freeman, 1977

Variable	Functional	Mixed pain	Mixed relief	Organic
Somatization	148.8	240.3	36.0	240.5
Low morale	148.8	81.0	116.6	176.9
Depression	256.0	84.6	49.0	34.8
Distortion	123.2	161.3	75.7	96.1

Meichenbaum, 1975

Creativity variable	Self-instruction	Focusing	Control
Fluency	26.01	5.29	18.49
Remote	8.41	4.41	.84
Obvious	10.89	10.89	5.76

[a] Groups are four levels of parent's education.
[b] Groups are four instructional sets.
[c] Groups are men classified according to degree of obesity.
[d] Groups formed on the basis of two types of contextual organization.
[e] Dependent variables are the amount of dental calculus on six anterior mandibular teeth.

Since V is only much more robust for extreme subgroup variance differences, the following question arises: "How often can one expect to find subgroup variances that differ by a magnitude of 36 to 1?" Table 2 presents the subgroup variances for a small sampling of studies from the literature. Although the sampling is small, it includes nine different sources and considers a variety of criterion variables (achievement, personality, biochemical, and dental). The first point these studies demonstrate is that subgroup variance differences near 36 to 1 occur very infrequently. Only in the Finn (1974) example are there differences of this magnitude, and even there the large differences are confined to certain variables and are not of the I type. The second point these studies illustrate is that the I type of heteroscedasticity Olson considered (in which one group has large variances on all variables and all other groups have equal and smaller variances) is one that probably does not occur very often in practice. Rather, the groups differ in a variety of ways. For example, in the French, Brownell, Graziano, and Hartup (1977) and Smith, Gnanadesikan, and Hughes (1962) studies the largest variances for the three variables in each case fell in three different groups. In Meichenbaum's (1975) study, although the self-instruction group had the largest variances on all three variables, the variances of the variables in the other two groups were far from equal.

To further document that extreme subgroup variance differences (especially of the 36I type) occur rarely in educational and psychological research, a check was made of three major journals: *American Educational Research Journal* for 1972–1977 and *Journal of Educational Psychology* and *Psychological Reports* for 1975–1977. This more extensive sampling confirmed the results of the smaller sampling reported in Table 2.

A secondary part of Olson's argument concerned the power of the four statistics. Which is most powerful depends on how the null hypothesis is false. With a concentrated noncentrality structure, the power ordering (from most to least powerful) is R, T, W, and V. With a diffuse structure (the groups differing along several dimensions), the power ordering is reversed (V, W, T, and R), and power differences among V, W, and T are typically small (Olson, 1973, p. 73).

Therefore, if assumptions are met, the choice of test statistic depends on the degree of concentration of the noncentrality structure. If assumptions are not met, then I would argue as follows: For concentrated or nearly concentrated structures, use V, W, or T as the test statistic, because the discrepancies in actual Type I error for these three, for subgroup variance differences likely to occur in practice, are so small that which of the statistics is used makes no practical difference. Although V will generally be slightly more robust, the gain in power obtained by using T or W instead of V will offset or more than offset V's slight robustness advantage. There is evidence to suggest that concentrated structures are quite prevalent in psychological research (Bock, 1975, p. 154). For diffuse structures, however, it must be acknowedged that V is the preferred choice, since in these cases it is slightly more robust and somewhat more powerful.

References

Bock, R. D. *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill, 1975.

Calsyn, D. A., Spengler, D. M., & Freeman, C. W. Application of the somatization factor of the MMPI-168 with low back pain patients. *Journal of Clinical Psychology,* 1977, *33,* 1017–1020.

Finn, J. D. *A general model for multivariate analysis.* New York: Holt, Rinehart & Winston, 1974.

French, D. C., Brownell, C. A., Graziano, W. G., & Hartup, W. W. Effects of cooperative, competitive and individualistic sets on performance in children's groups. *Journal of Experimental Child Psychology,* 1977, *24,* 1–10.

Gardner, E. T., & Schumacher, G. M. Effects of contextual organization on prose retention. *Journal of Educational Psychology,* 1977, *69,* 146–151.

Meichenbaum, D. Enhancing creativity by modifying what subjects say to themselves. *American Educational Research Journal,* 1975, *12,* 129–145.

Novince, L. *The contribution of cognitive restructuring to the effectiveness of behavior rehearsal in modifying social inhibition in females.* Unpublished doctoral dissertation, University of Cincinnati, 1977.

Olson, C. L. *A Monte Carlo investigation of the robustness of multivariate analysis of variance.* Unpublished doctoral dissertation, University of Toronto, 1973.

Olson, C. L. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 1976, *83*, 579–586.

Smith, H., Gnanadesikan, R., & Hughes, J. B. Multivariate analysis of variance. *Biometrics*, 1962, *18*, 22–41.

Stevens, J. P. Four methods of analyzing between variation for the k-group MANOVA problem. *Multivariate Behavioral Research*, 1972, *7*, 499–522.

Wright, R. J. The affective and cognitive consequences of an open education elementary school. *American Educational Research Journal*, 1975, *12*, 449–468.

Received December 12, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 361–370

Confirmatory Inference and Geometric Models

Lawrence J. Hubert
Department of Education
University of California, Santa Barbara

Michael J. Subkoviak
University of Wisconsin—Madison

A confirmatory technique is discussed that is appropriate for comparing a given geometric model with supplementary data available on the same objects used in the representation. The inference procedure is based on relatively straightforward distribution-free principles and requires the comparison of one proximity matrix, possibly reconstructed from a particular geometric model, with a second structure matrix obtained from the supplementary information. Two examples are presented that illustrate the generality of the same statistical approach.

In the literature on data analysis over the last 20 years, a distinction between exploratory and confirmatory procedures has become very popular (Hildebrand, Laing, & Rosenthal, 1977; Kaiser, 1970; Tukey, 1962). An exploratory strategy typically involves the use of an analysis technique on a given data set with the aim of identifying interesting relationships, patterns, and the like. Alternatively, a confirmatory approach requires the test of an a priori conjecture that is generated from a source distinct from the data to be used for the purposes of validation. This latter test in the present context is correlational, and thus the term *confirmation* is given a limited meaning that does not imply the absolute correctness of a hypothesis. Since a correlational analysis can never exclude all competing explanations, we argue when it is justified that the pattern of data is not unrelated to the conjectured pattern.

It may be obvious that confirmatory analyses would be desirable adjuncts to many of the current exploratory methods used in the study of proximity matrices (such as clustering and multidimensional scaling), but very few techniques have been proposed that could help carry out such a program with any degree of rigor. Users of the newer data reduction procedures lack confirmatory techniques even of a correlational nature and must rely on intuitive arguments based on whatever additional information is available for the objects being studied. Although this practice is commendable given the current state of the art, it is now possible to proceed one step further using the correlational methods presented in this article and incorporate the same information relevant to a post hoc explanation more directly in a confirmatory manner. To provide a complete illustration and also to limit the scope of the discussion, our emphasis is on geometric models, or more generally, on data representations that can be given some type of explicit geometric interpretation. In short, the sections to follow illustrate how confirmatory data analysis problems that are phrased geometrically can be approached through a relatively straightforward concept of correlation. Depending on the context, it is conceivable that the method presented here might be used in conjunction with a geometric model; in extending an existing analysis strategy based on geometric notions; as a means of interpreting a given model with respect to supplementary information; or finally, as a preliminary to the construction of a desired geometric representation. The first example

---

Equal authorship is implied. Lawrence J. Hubert and Michael J. Subkoviak were supported respectively by National Science Foundation Grant SOC 75-07860 and National Institute of Education Grant NIE-6-75-0088. The authors wish to thank Phipps Arabie and several reviewers for their helpful comments and Myron Wish for the data used in Example 2.

Requests for reprints should be sent to Lawrence J. Hubert, Department of Education, University of California, Santa Barbara, California 93106.

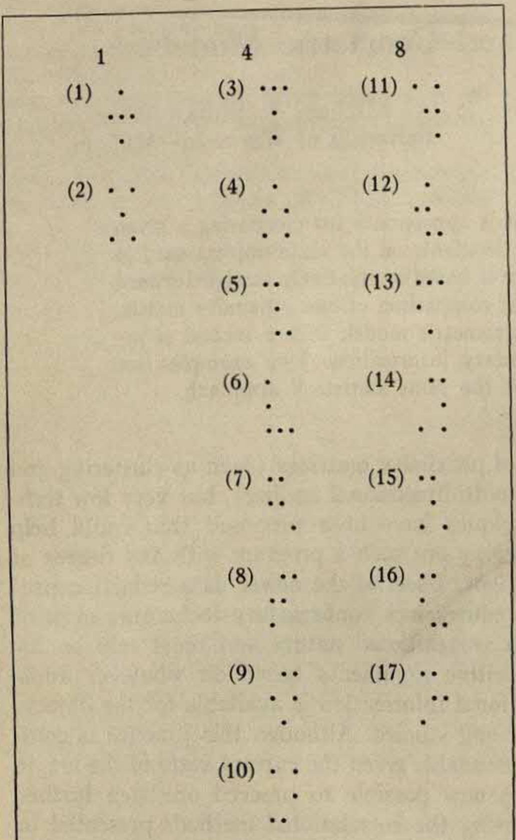

Figure 1. Patterns used by Glushko (1975) in testing Garner's (1962) pattern goodness hypothesis; each pattern is characterized by one of three inferred equivalence class sizes: 1 and 2 by a size of one, 3–10 by a size of four, and 11–17 by a size of eight. Glushko's patterns were as wide as they were high.

below formalizes the basic ideas that are presented here.

Example 1: Figural Goodness

In a recent study concerned with the "goodness" of patterns, Glushko (1975) attempted to verify Garner's (1962) basic hypothesis regarding what makes one pattern better than another. To be more specific, each of the 17 patterns used by Glushko, listed in Figure 1, can be characterized by the size of an inferred equivalence class. The term *equivalence* is used to label the set of patterns that contain a single figure plus all other configurations that result from reflections or from 90° rigid rotations. As indicated in Figure 1, 2 of the Glushko patterns construct the same configuration under all of these operations, 8 patterns have four associated figures, and finally, 7 patterns produce 8 different members. According to Garner, the subjective judgment of pattern goodness is a direct function of the size of a configuration's inferred equivalence class, with the smaller size classes corresponding to the better patterns.

To test Garner's hypothesis using the 17 patterns of Figure 1, Glushko first obtained a symmetric measure of proximity between each pair of patterns by using a choice task. All 136 different pattern combinations were presented to 20 subjects, who were asked to indicate their preference. These choices were then summed over subjects and subtracted from an expected preference frequency of 10. Due to the subtraction of 10, the absolute values of these differences, given in the lower triangular portion of Table 1, form a symmetric measure of proximity defined for all pattern pairs and provide data in a form that can be subjected to a variety of data reduction techniques. In particular, Glushko attempted to represent the structure of the proximity function by first placing the 17 configurations in a two-dimensional space using Shepard and Kruskal's multidimensional scaling routine (see Kruskal, 1964a, 1964b). Given this geometric representation, Johnson's (1967) diameter clustering results were then superimposed, producing a representation similar to that we give in Figure 2 (here, we only indicate the clustering result defined by three subclasses). Clearly, one strong dimension (the vertical) can be identified as that of equivalence class size. In addition, the clusters themselves correspond fairly well to a grouping on the basis of the same criterion except for the minor misplacement of the two configurations numbered 10 and 11.

The process of verifying Garner's hypothesis through a multidimensional scaling and clustering might be considered rather circuitous, especially since the equivalence class hypothesis implies a definite structure for the original proximity measure. Although one dimension is very strong in this example and the clustering and scaling results are clear-cut, unambiguous outcomes of this type are

Table 1
Symmetric Proximity Matrix Obtained by Glushko (1975) for the Patterns of Figure 1 and Structure Matrix Generated by Equivalence Class Hypothesis[a]

Pattern	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	X	0	3	3	3	3	3	3	3	3	7	7	7	7	7	7	7
2	1	X	3	3	3	3	3	3	3	3	7	7	7	7	7	7	7
3	1	2	X	0	0	0	0	0	0	0	4	4	4	4	4	4	4
4	2	4	0	X	0	0	0	0	0	0	4	4	4	4	4	4	4
5	3	3	1	1	X	0	0	0	0	0	4	4	4	4	4	4	4
6	2	4	1	1	1	X	0	0	0	0	4	4	4	4	4	4	4
7	2	4	3	2	1	2	X	0	0	0	4	4	4	4	4	4	4
8	3	5	2	1	2	1	0	X	0	0	4	4	4	4	4	4	4
9	4	4	2	1	5	3	3	4	X	0	4	4	4	4	4	4	4
10	4	5	4	4	3	3	3	5	4	X	4	4	4	4	4	4	4
11	5	5	3	4	3	0	2	3	1	1	X	0	0	0	0	0	0
12	5	6	4	6	4	1	5	5	2	1	3	X	0	0	0	0	0
13	6	7	7	6	5	4	5	6	5	1	4	1	X	0	0	0	0
14	7	6	4	4	5	4	6	5	4	2	4	1	1	X	0	0	0
15	6	7	5	7	4	5	5	4	5	0	3	0	0	1	X	0	0
16	7	8	5	5	6	4	4	3	4	1	4	2	2	0	1	X	0
17	7	7	5	5	5	6	5	4	6	3	6	2	3	1	1	1	X

[a] Symmetric proximity matrix is in lower triangle; structure matrix is in upper triangle.

somewhat rare. In general, when a strong hypothesis is not reflected as dramatically in the scaling or clustering results, it may be difficult to decide whether the hypothesis is inadequate or whether the data reduction techniques are at fault. In the typical application, the researcher may be able to identify portions of his or her theory in a scaling or clustering solution, but may lack a strategy for measuring in any precise manner the actual degree of confirmation or nonconfirmation.

As an alternative approach, it should be possible to test directly whether the pattern goodness hypothesis is reflected in the original proximities and bypass the scaling and clustering solutions altogether. To introduce some notation, suppose the patterns are denoted by $o_1, o_2, \ldots, o_n$ (where $n = 17$ in our example). Furthermore, let $q(o_i,o_j)$ refer to the symmetric[1] proximity between patterns o_i and o_j, and let **Q** refer to an organization of these measures into a 17×17 square matrix with rows and columns labeled by the objects or patterns $o_1, o_2, \ldots, o_n$. By convention, the diagonal of **Q** is assumed to consist entirely of zeros. In addition to the empirical proximity matrix **Q**, the stated hypothesis is represented numerically by a second "structure" matrix **C** with elements $c(o_i,o_j)$. Explicitly, suppose $N(o_i)$ denotes the size of the inferred equivalence class for object o_i, and let f be some monotone function on the integers, for example, $f(x) > f(y)$ if and only if $x > y$. Then as a formal definition, $c(o_i,o_j) = f[|N(o_i) - N(o_j)|]$, where it is assumed that $c(o_i,o_j) = 0$ for $o_i = o_j$. Although many functions can be used and the actual choice depends on the researcher's judgment as to the most appropriate relative size of the structure values, for the purposes of an illustration f is taken as the identity, that is, $f(x) = x$. In other words, the symmetric function values $c(o_i,o_j)$, given in the upper triangular portion of Table 1, are merely the absolute values of the differences in equivalence class sizes associated with the objects o_i and o_j.

As an operational interpretation, the theory used to generate the function $c(o_i,o_j)$ is given empirical support if the two sets of elements, $c(o_i,o_j)$ and $q(o_i,o_j)$, have a similar patterning of high and low entries. Although many formal indices for this relationship can be defined, the pairing of a proximity $q(o_i,o_j)$ with a structure value $c(o_i,o_j)$ suggests that the

[1] Extensions to asymmetric measures are possible using the generalizations discussed in Hubert and Schultz (1976).

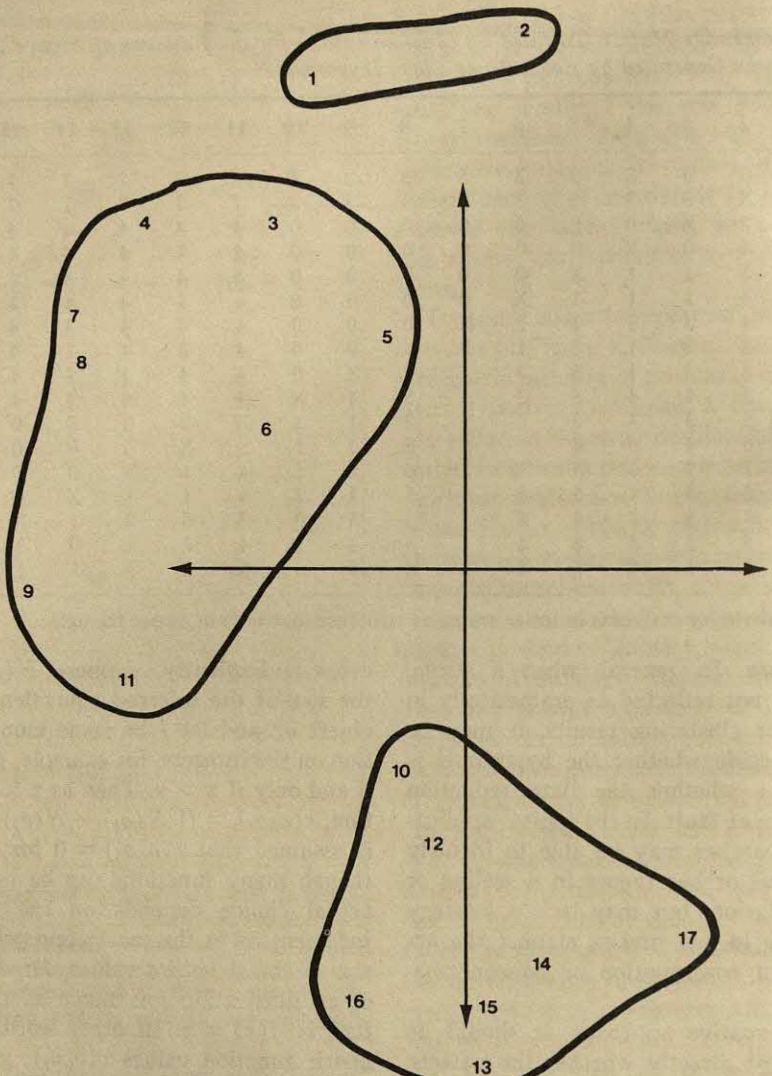

Figure 2. Two-dimensional scaling of the 17 Glushko (1975) patterns.

simple Pearson product-moment correlation may be a reasonable measure to consider; it is thus our major choice for the sequel.[2] Once this index is calculated, the next problem concerns its significance and, specifically, whether the size of the observed correlation between the values for $q(o_i,o_j)$ and $c(o_i,o_j)$ is sufficient to reject some appropriately defined null hypothesis.

To generate a reasonable reference distribution for the observed correlation, suppose one assumes a randomness hypothesis that can hopefully be rejected. More specifically, it is conjectured that the partition of the objects (or patterns) $o_1, o_2, \ldots, o_n$ occurred randomly or was chosen at random from the set of all

[2] Since the conditional assumption of fixed functions $c(o_i,o_j)$ and $q(o_i,o_j)$ is made, several other indices are formally equivalent to the Pearson statistic, at least from an inference point of view. For instance, either the raw sum of cross-products, $\Sigma_{i,j}\ c(o_i,o_j)q(o_i,o_j)$, or the sum of squared differences, $\Sigma_{i,j}\ [c(o_i,o_j) - q(o_i,o_j)]^2$, could be used. Alternatively, a transformation on the original function values could be carried out and the comparison performed on the transformed values. Thus, if the ranks of $c(o_i,o_j)$ and $q(o_i,o_j)$ are used, then Spearman's rank correlation index is the measure of correspondence.

partitions of the same form. In our case, the conjectured partition contains three classes with two, eight, and seven objects in each, and thus, the null hypothesis of interest asserts that this particular partition occurred randomly and consequently does not reflect the patterning of entries in the proximity matrix Q. Moreover, any such partition of Q of the same form (i.e., number of classes) will produce a correlation index and when completely enumerated will generate an exact reference distribution for the null hypothesis. Since it is assumed that the partition is conjectured prior to observing the data, the hypothesis of randomness is an active candidate to consider, and its rejection is not a foregone conclusion.

From an inference perspective, the observed correlation for the conjectured partition can be compared with the distribution generated by complete enumeration, and if the observed correlation is at a suitably extreme percentage point, the null hypothesis of randomness can be rejected. Moreover, when the correlation actually obtained for the conjectured partition is large enough, this index can be assumed to reflect a value that was obtained nonrandomly; that is, at least to some extent, the functions $q(o_i,o_j)$ and $c(o_i,o_j)$ have a common patterning of high and low entries. In short, a permutation (or randomization) test of the type discussed in detail by Bradley (1968), Edgington (1969), or Lehmann (1975) is applied. For instance, if the functions $q(o_i,o_j)$ and $c(o_i,o_j)$ are specialized appropriately (Hubert & Baker, 1977), then this same strategy includes the usual randomization analysis of a one-way design.

Although complete enumeration is generally prohibitive because of computational costs and although an exact reference distribution is thus typically too expensive to obtain, Monte Carlo approximations are relatively inexpensive (cf. Hubert & Schultz, 1976; Schultz & Hubert, 1976). For instance, Table 2 presents the frequency results of selecting 1,000 partitions of the desired form at random and with replacement and should provide an approximate distribution that is fairly accurate for this application. In particular, using the Figure 1 data, the observed correlation for the Garner (1962) hypothesis is .64, which

Table 2
Approximate Distribution for Comparison of Structure and Proximity Matrices in Table 1

Correlation	Sample cumulative proportion
−.193	.001
−.171	.005
−.162	.010
−.117	.050
−.098	.100
−.070	.200
−.046	.300
−.025	.400
−.009	.500
.010	.600
.033	.700
.068	.800
.115	.900
.162	.950
.273	.990
.297	.995
.396	.999
.420	1.000

Note. $N = 1{,}000$.

is greater than any value observed in the Table 2 distribution. Thus, the null hypothesis of a random partition can be rejected at an approximate significance level of, say, .001, suggesting that the equivalence class hypothesis is supported by the patterning of the proximity values.

Using the previous example as a guide, the salient features of a confirmatory analysis should be evident. Given the proximity measure $q(o_i,o_j)$ and some conjecture specified in terms of a structure function $c(o_i,o_j)$, the observed correlation between $q(o_i,o_j)$ *and* $c(o_i,o_j)$ is compared with a reference distribution generated under a hypothesis of randomness. If the obtained correlation is at an extreme percentage point, the correspondence between $q(o_i,o_j)$ and $c(o_i,o_j)$ is declared significant, with the added implication that the conjecture leading to the construction of $c(o_i,o_j)$ may help explain some of the variation present in the empirical proximity measures.[3] As

[3] If different functions $c(o_i,o_j)$ are used on the same data set, our confirmatory strategy (as well as any other) could turn out to be an exploratory analysis. In this case, however, the difficulty of multiple significance testing arises.

usual, the size of the correlation can be considered an index of the degree of correspondence or confirmation.

Although the example given above implies that a randomness hypothesis should be defined in terms of selecting a partition of a given form at random, a more general hypothesis that will generate exactly the same distribution can also be considered. Explicitly, if the values assigned by the proximity function are organized, as before, into an $n \times n$ square matrix $\mathbf{Q}$ and, similarly, the values of the structure function into a second $n \times n$ square matrix, $\mathbf{C}$, both with rows and columns labeled as $o_1, o_2, \ldots, o_n$, then each reordering of the rows and simultaneously of the corresponding columns of $\mathbf{Q}$ in relation to the fixed $\mathbf{C}$ matrix will induce a specific partition of the n objects $o_1, \ldots, o_n$. In other words, for our $\mathbf{C}$ matrix of Table 1, any reordering of $\mathbf{Q}$ produces a partition defined by subsets containing two, eight, and seven objects. The first two rows and columns of the reordered $\mathbf{Q}$ matrix define the objects in the class of size two, the next eight rows and columns define a class of eight objects, and the remaining seven rows and columns define the last object class of size seven. Morover, if a reordering of $\mathbf{Q}$ is chosen at random, that is, if all $n!$ possible reorderings are considered equally likely, then this assumption induces a random selection of a partition of the same general form used in the original construction of the $\mathbf{C}$ matrix. In short, the random reordering of $\mathbf{Q}$ and the random selection of a partition will generate exactly the same distribution of correlations, and thus, either concept can be used in producing an approximate reference table through Monte Carlo simulation. This generalization will prove important when a confirmatory approach is necessary but cannot be identified by a specific partitioning of an object set.

Although we suggest carrying out a confirmatory test through the use of an approximate distribution obtained through Monte Carlo simulation, it is also possible to find the exact mean and variance of the complete reference distribution by formulas, given only the matrices $\mathbf{C}$ and $\mathbf{Q}$. Specifically, the mean of the Pearson correlation r is zero, and its variance is equal to

$$V(r) = [1/\tbinom{n}{2}]\left\{1 + \frac{1}{(n-2)G_2H_2} \times \left[2(G_1 - G_2)(H_1 - H_2) + \frac{(2G_1 - G_2)(2H_1 - H_2)}{(n-3)}\right]\right\},$$

where

$$G_1 = \sum_i \{\sum_{\substack{j \\ i \neq j}} [q(o_i,o_j) - \bar{q}]\}^2;$$

$$G_2 = \sum_i \sum_{\substack{j \\ i \neq j}} [q(o_i,o_j) - \bar{q}]^2;$$

$$H_1 = \sum_i \{\sum_{\substack{j \\ i \neq j}} [c(o_i,o_j) - \bar{c}]\}^2;$$

$$H_2 = \sum_i \sum_{\substack{j \\ i \neq j}} [c(o_i,o_j) - \bar{c}]^2;$$

$$\bar{q} = \frac{2}{n(n-1)} \sum_{i<j}\sum q(o_i,o_j);$$

$$\bar{c} = \frac{2}{n(n-1)} \sum_{i<j}\sum c(o_i,o_j).$$

As an example of how the variance calculation can be used for the data of Figure 1 and the structure function of Table 1, we find $[V(r)]^{\frac{1}{2}} = .0879$. Converting to a Z score for the observed correlation of .641, a value of 7.28 is obtained, which would indicate a rather significant result if it were possible to assume even a crude normality (see Mantel, 1967, for the appropriate moment derivations).[4]

Although the Pearson product-moment correlation is used as a measure of correspondence between $c(o_i,o_j)$ and $q(o_i,o_j)$, it is not legitimate to test the index using the usual

[4] Unfortunately, since the few normal convergence theorems that are available are also very specialized, little general information is available on the adequacy of a normal approximation. Several Monte Carlo studies, however, suggest that normality may provide an adequate approximation in some applications (e.g., Schultz & Hubert, 1976). Thus, until more complete information is available, it may be more appropriate to rely on random sampling from the complete permutation distribution or to use the exact moments to obtain a conservative significance level, as discussed in Hubert and Levin (1976a).

formulas presented in most elementary tests. In particular, the permutation procedure discussed above preserves the internal linkages among the function values $c(o_i,o_j)$ and $q(o_i,o_j)$, whereas an application of even the usual permutation test on a correlation, as discussed in Bradley (1968), does not. In the latter case, all $\binom{n}{2}!$ reorderings of one set of function values would be considered equally likely and would be compared with some fixed ordering of the second set of $\binom{n}{2}$ function values. Thus, dependencies among the function values would be destroyed and a different variance of $1/[\binom{n}{2} - 1]$ for r would result. These same comments apply to the use of the well-known parametric hypothesis test of no correlation based on the t distribution.

Example 2: Multidimensional Scaling Applications

Geometric configurations that are generated as a result of a multidimensional scaling (Carroll & Chang, 1970; Kruskal, 1964a, 1964b) represent another context in which the confirmatory paradigm could be used to test a priori conjectures. To give an illustration, consider the application of Carroll and Chang's individual differences scaling procedure to data collected by Wish, Deutsch, and Biener (1970). The objects of study for this analysis were 12 nations, and each of 18 subjects rated the proximity of all pairs of nations on a 9-point scale (large numbers indicated a greater degree of similarity). The resulting 18 proximity matrices, all of size 12×12, can be analyzed by the Carroll–Chang procedure. The group result selected for our discussion is a two-dimensional configuration, shown in Figure 3, in which each nation is represented by a point and in which the interpoint distances reflect the degree of similarity between the corresponding nations as judged by the group; for example, the distance between the United States and China is large, since they are perceived on the average as being very dissimilar. (It should be noted at the outset that two separate two-way analyses are performed below, and the interlocking

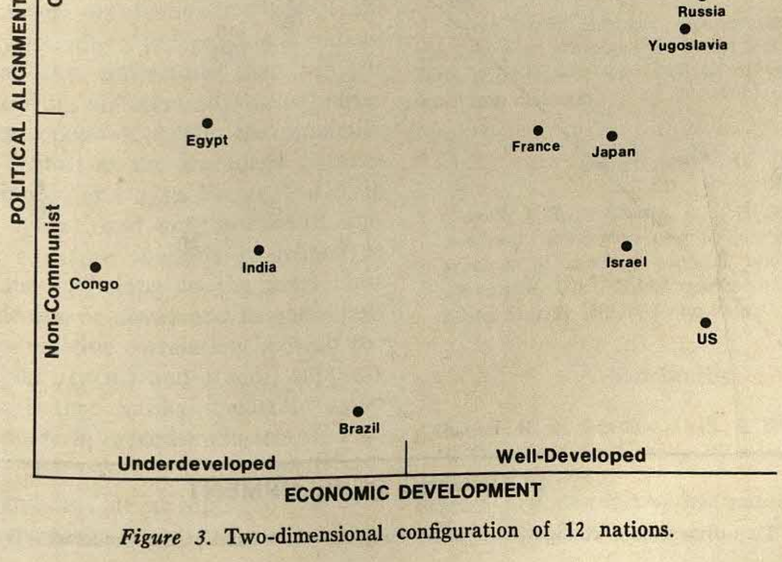

Figure 3. Two-dimensional configuration of 12 nations.

weights between what are called the group and subject spaces are not considered.)

Instead of attempting to label and interpret dimensions per se, suppose the researcher wishes to test the a priori hypothesis that an outside variable, such as political alignment, may account in part for the distances between nations. In other words, the researcher is interested in confirming the conjecture that nations close together subscribe to similar political philosophies and, conversely, that those far apart have different political systems. In this case the proximity function $q(o_i,o_j)$ would merely refer to the distance between nations o_i and o_j in Figure 3.

The structure function $c(o_i,o_j)$ would be obtained from the outside variable of political alignment. For instance, if political alignment were simply dichotomized as communist versus noncommunist, then $c(o_i,o_j)$ might be defined as zero if o_i and o_j were both communist or both noncommunist and as one otherwise. With this notation, a large positive correlation between $q(o_i,o_j)$ and $c(o_i,o_j)$ would indicate that nations of similar political persuasion were located close together in Figure 3. As it turns out, the observed correlation between the interpoint distances in Figure 3 and the dichotomous variable of political alignment is .50, which is significant at, say, the .001 level (approximately) when referred to the distribution of correlations for 1,000 random reorderings of matrix Q. In short, there is statistical support for the hy-

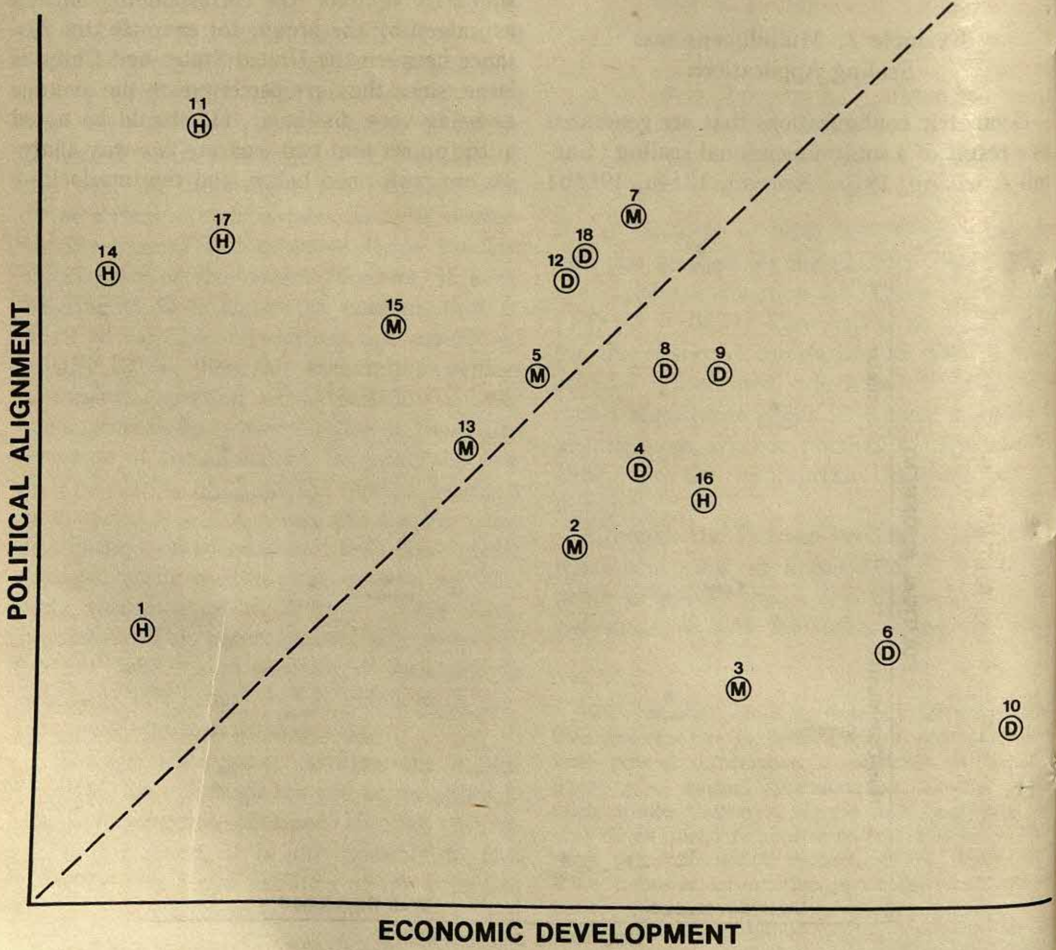

Figure 4. Two-dimensional configuration of 18 subjects. (H = hawk; M = moderate; D = dove.)

pothesis that the information provided by the variable of political alignment (or some closely related variable) is reflected in the arrangement of points. The question regarding other possible competing hypotheses, however, is unanswered.

In addition to the geometric configuration of nations given in Figure 3, the Carroll–Chang procedure also produces a configuration of the particular subjects that supplied the similarity data, as shown in Figure 4. The horizontal and vertical axes of Figure 4 are exactly the same as those of Figure 3, representing economic development and political alignment, respectively. Numerical coordinates (w_{i1}, w_{i2}) again locate a subject o_i in Figure 3 and, furthermore, indicate how much emphasis subject o_i gives to political alignment and economic development when rating the similarities of nations. Thus, subject 10 gives primary emphasis to the economic development dimension, subject 11 gives primary emphasis to political alignment, and subjects in the center of the configuration weight both dimensions about equally.

As indicated in Figure 4, Wish et al. further classified each subject either as a hawk (H), a moderate (M), or a dove (D) according to the person's stance on the Vietnam War and descriptively argued that subjects in the same class tend to weight the two dimensions similarly. Since it is hypothesized that hawks, moderates, and doves will form reasonably homogeneous clusters in Figure 4, the confirmatory paradigm can provide a statistical test for the conjecture that subjects weight dimensions differentially according to their political opinions. Again, the proximity function is defined as the Euclidean distance between points o_i and o_j in Figure 4 (i.e., in the subject space), and for the sake of simplicity, the structure function is defined as zero if o_i and o_j belong to the same class (hawk, moderate, or dove) and as one otherwise. A large positive correlation between the function values $q(o_i,o_j)$ and $c(o_i,o_j)$ supports the conjecture that hawks, moderates, and doves tend to form separate clusters in Figure 4. Since the observed correlation is .19, which is significant at an approximate .009 level, the hypothesis is given statistical support. Wish et al. noted specifically that hawks tend to cluster above the diagonal in Figure 4 and give relatively more emphasis to the political alignment factor, whereas moderates and doves cluster below the diagonal and give relatively more weight to economic development.[5]

Discussion

As should be evident in the examples given above, the confirmatory approach developed in this article has a number of applications that are related to the use and development of geometric models, either those that occur naturally or those derived from some intermediate data reduction process. In addition to the illustrations provided, a number of other correspondences to the methodological literature of the behavioral sciences could be developed that the reader may be interested in pursuing further; see, for example, Schultz and Hubert (1976), Hubert and Baker (1977), Hubert and Levin (1976a, 1976b), Hubert and Schultz (1976), Hubert (1978), Carroll and Chang (Note 1), Althauser, Burdick, and Winsborough (1966), Campbell, Kruskal, and Wallace (1966), Cliff & Ord (1973), Geary (1954), Mielke, Berry, and Johnson (1976), Royaltey, Astrachan, and Sokal (1975), and Winsborough, Quarantelli, and Yutzy (1963).

---

[5] Since strong algebraic dependencies exist among the interpoint distances, it is not legitimate to perform a simple one-way analysis of variance on the Euclidean distances.

Reference Note

1. Carroll, J. P., & Chang, J. J. *A general index of nonlinear correlation and its application to the problem of relating physical and psychological dimensions.* Unpublished manuscript, Bell Laboratories, Murray Hill, N.J., no date.

References

Althauser, R. P., Burdick, D. S., & Winsborough, H. H. The standardized contiguity ratio. *Social Forces*, 1966, *45*, 237–245.

Bradley, J. V. *Distribution-free statistical tests.* New York: Wiley, 1968.

Campbell, D. T., Kruskal, W. H., & Wallace, W. P. Seating aggregation as an index of attitude. *Sociometry*, 1966, *29*, 1–15.

Carroll, J. D., & Chang, J. J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 1970, *35*, 283–319.

Cliff, A. D., & Ord, J. K. *Spatial autocorrelation*. London: Pion, 1973.

Edgington, E. S. *Statistical inference: The distribution-free approach*. New York: McGraw-Hill, 1969.

Garner, W. R. *Uncertainty and structure as psychological concepts*. New York: Wiley, 1962.

Geary, R. C. The contiguity ratio and statistical mapping. *Incorporated Statistician*, 1954, *5*, 115–141.

Glushko, R. J. Pattern goodness and redundancy revisited: Multidimensional scaling and hierarchical clustering analyses. *Perception & Psychophysics*, 1975, *17*, 158–162.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. *Prediction analysis of cross-classifications*. New York: Wiley, 1977.

Hubert, L. J. Nonparametric tests for patterns in geographic variation: Possible generalizations. *Geographical Analysis*, 1978, *10*, 86–88.

Hubert, L. J., & Baker, F. B. Analyzing distinctive features. *Journal of Educational Statistics*, 1977, *2*, 79–98.

Hubert, L. J., & Levin, J. R. Evaluating object set partitions: Free-sort analysis and some generalizations. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 459–470. (a)

Hubert, L. J., & Levin, J. R. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 1976, *83*, 1072–1080. (b)

Hubert, L. J., & Schultz, J. V. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 1976, *29*, 190–241.

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, *32*, 241–254.

Kaiser, H. F. A second generation little jiffy. *Psychometrika*, 1970, *35*, 401–415.

Kruskal, J. B. Multidimensional scaling: A numerical method. *Psychometrika*, 1964, *29*, 1–27. (a)

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, *29*, 115–129. (b)

Lehmann, E. L. *Nonparametrics*. San Francisco: Holden-Day, 1975.

Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 1967, *27*, 209–220.

Mielke, P. W., Berry, K. J., & Johnson, E. S. Multiresponse permutation procedures for a priori classifications. *Communications in Statistics. Part A: Theory and Methods*, 1976, *5*, 1409–1424.

Royaltey, H. H., Astrachan, E., & Sokal, R. R. Tests for patterns in geographic variation. *Geographical Analysis*, 1975, *7*, 369–395.

Schultz, J. V., & Hubert, L. J. A nonparametric test for the correspondence between two proximity matrices. *Journal of Educational Statistics*, 1976, *1*, 59–67.

Tukey, J. W. The future of data analysis. *Annals of Mathematical Statistics*, 1962, *33*, 1–67.

Winsborough, H. H., Quarantelli, E. L., & Yutzy, D. The similarity of connected observations. *American Sociological Review*, 1963, *28*, 977–983.

Wish, M., Deutsch, M., & Biener, L. Differences in conceptual structures of nations: An exploratory study. *Journal of Personality and Social Psychology*, 1970, *16*, 361–373.

Received December 16, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 371–375

Nonparametric Large-Sample Pairwise Comparisons

Kenneth J. Levy
State University of New York at Buffalo

Tukey's procedure for making pairwise comparisons among means is discussed within the context of three nonparametric models. Examples are presented in which Tukey's procedure, in accord with Hartley's results, is employed to make comparisons associated with a Kruskal-Wallis one-way analysis of variance test for ranked data, Friedman's two-way analysis of variance test for ranked data, and Cochran's test of change for dichotomous data.

Marascuilo and McSweeney (1967) discussed methods for testing post hoc hypotheses concerning trends associated with the Kruskal–Wallis (1952) one-way analysis of variance (ANOVA) test for ranked data, the Friedman (1937) two-way ANOVA test for ranked data, and the Cochran (1950) test of change for dichotomous data.

Within the framework of these three nonparametric tests, an investigator might be primarily interested in testing hypotheses associated with the set of all possible pairwise comparisons that arise from the k treatments associated with each of these three nonparametric test procedures. The purpose of the present article is to illustrate the application of a Tukey-type procedure for controlling the joint significance level associated with such comparisons. Marascuilo and McSweeney's general multiple comparison approach could be employed in the present context if an investigator were only interested in making pairwise comparisons; however, the present procedure produces more powerful tests because the set of pairwise comparisons is only a subset of the set of all possible comparisons for which Marascuilo and McSweeney's approach is most applicable.

The Kruskal–Wallis Test

Following Marascuilo and McSweeney (1967), consider k independent samples, each of size n, drawn from k continuous probability distributions. Let $N = nk$ and let the original observations be replaced by ranks $(1, 2, \ldots, N)$ in accord with the Kruskal–Wallis test. Let $R_1, R_2, \ldots, R_k$ be the rank sums of each of the samples, and let $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ be the respective average ranks.

From Marascuilo and McSweeney, when $E(\bar{R}_1) = E(\bar{R}_2) = \cdots = E(\bar{R}_k)$, where $E(\ )$ is the expected value operator, it should be noted that if n is sufficiently large, the $\bar{R}_j$ will be approximately multivariate normal with

$$E(\bar{R}_j) = \frac{(N+1)}{2},$$

$$\operatorname{var}(\bar{R}_j) = \sigma_j^2 = \frac{(N+1)(N-n)}{12n},$$

and

$$\operatorname{cov}(\bar{R}_j, \bar{R}_{j'}) = \sigma_{jj'} = \frac{-(N+1)}{12}$$

$$= \sigma_j^2\left[\frac{-n}{(N-n)}\right].$$

Consider the following result from Hartley (1950). If $\mathbf{y}' = (y_1, y_2, \ldots, y_k)$ is a multivariate normal vector with zero means and dispersion $\sigma^2\mathbf{B}$ (where $\mathbf{B}$ is a square matrix whose diagonal elements equal a and whose off-diagonal elements equal b for some a and b), then the range of the y_js is distributed as the range of k independent and identically distributed normal variates with zero means and variances $\sigma^2(a-b)$. In the present context, in accord with Hartley's result, the range of the $\bar{R}_j$s is distributed as the range of k in-

Requests for reprints should be sent to Kenneth J. Levy, Department of Psychology, State University of New York, 4230 Ridge Lea Road, Buffalo, New York 14226.

dependent and identically distributed normal variates with zero means and variances $\sigma^2(a-b)$, where $\sigma^2(a-b) = N(N+1)/12n$. So

$$P\{(\bar{R}_{\max} - \bar{R}_{\min}) \le q_{\alpha,k,\infty}[N(N+1)/12n]^{\frac{1}{2}}\} = 1 - \alpha,$$

where $\bar{R}_{\max} - \bar{R}_{\min}$ is the range of the $\bar{R}_j$s and $q_{\alpha,k,\infty}$ is the upper α point of the studentized range with infinite degrees of freedom. From this probability statement follows the usual Tukey result that the probability is $1-\alpha$ that all of the $(1/2)k(k-1)$ pairwise differences $[E(\bar{R}_j) - E(\bar{R}_{j'})]$ simultaneously satisfy

$$(\bar{R}_j - \bar{R}_{j'}) - \omega q_{\alpha,k,\infty} \le [E(\bar{R}_j) - E(\bar{R}_{j'})] \le (\bar{R}_j - \bar{R}_{j'}) + \omega q_{\alpha,k,\infty},$$

where $\omega = [N(N+1)/12n]^{\frac{1}{2}}$.

For data involving r sets of tied observations, let t_s be the number of tied observations in set s. Kruskal and Wallis (1952) suggested that for any group of t_s tied observations, the tied ranks could be replaced by their mean. From Dunn (1964, p. 249), when $E(\bar{R}_1) = E(\bar{R}_2) = \cdots = E(\bar{R}_k)$, it can be demonstrated that if n is sufficiently large, the $\bar{R}_j$ will be approximately multivariate normal with

$$E(\bar{R}_j) = \frac{(N+1)}{2},$$

$$\operatorname{var}(\bar{R}_j) = \left[\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r}(t_s^3 - t_s)}{12(N-1)}\right] \times \left(\frac{N-n}{nN}\right),$$

and

$$\operatorname{cov}(\bar{R}_j, \bar{R}_{j'}) = \left(\frac{-n}{N-n}\right)\operatorname{var}(\bar{R}_j).$$

Thus, in accord with Hartley's result, when ties occur in the data, the range of the $\bar{R}_j$s is distributed as the range of k independent and identically distributed normal variates with zero means and variances:

$$(\omega^*)^2 = \left(\frac{N}{N-n}\right)\operatorname{var}(\bar{R}_j).$$

Therefore, for tied data, it follows that the probability is $1-\alpha$ that all possible pairwise comparisons $[E(\bar{R}_j) - E(\bar{R}_{j'})]$ simultaneously satisfy

$$(\bar{R}_j - \bar{R}_{j'}) - \omega^* q_{\alpha,k,\infty} \le [E(\bar{R}_j) - E(\bar{R}_{j'})] \le (\bar{R}_j - \bar{R}_{j'}) + \omega^* q_{\alpha,k,\infty},$$

where

$$\omega^* = \left[\frac{N(N+1)}{12n} - \frac{\sum_{s=1}^{r}(t_s^3 - t_s)}{12n(N-1)}\right]^{\frac{1}{2}}.$$

A Kruskal–Wallis-Type Example

Consider the example discussed in Marascuilo and McSweeney (1967, p. 404). This example may also be found in Hays (1973, p. 684). The hypothetical example involves noise intensity as a treatment variable with six levels. The dependent variable is a subject's score obtained in a complex performance task under one of the noise intensity levels. The data for this example appear in Table 1.

An investigator wishes to test the 15 different hypotheses associated with pairwise comparisons of the form $[E(\bar{R}_j) - E(\bar{R}_{j'})] = 0$. Since there are $r = 19$ sets of tied observations, ω^* should be calculated. For these data, $\omega^* = 5.52$. If the joint significance level of the 15 tests is to be controlled at $\alpha = .05$, one also needs to obtain the value of $q_{.05,6,\infty}$. This value may be found in many statistical texts; and $q_{.05,6,\infty} = 4.03$. Those comparisons for which

$$|\bar{R}_j - \bar{R}_{j'}| > \omega^* q_{\alpha,k,\infty}$$

should be declared significant. In the present example, $\omega^* q_{\alpha,k,\infty} = 22.25$, and

$\lvert\bar{R}_1-\bar{R}_2\rvert = 17.75$	$\lvert\bar{R}_2-\bar{R}_3\rvert = 7.05$	$\lvert\bar{R}_3-\bar{R}_5\rvert = 33.00^*$
$\lvert\bar{R}_1-\bar{R}_3\rvert = 24.80^*$	$\lvert\bar{R}_2-\bar{R}_4\rvert = 8.70$	$\lvert\bar{R}_3-\bar{R}_6\rvert = 43.00^*$
$\lvert\bar{R}_1-\bar{R}_4\rvert = 9.05$	$\lvert\bar{R}_2-\bar{R}_5\rvert = 25.95^*$	$\lvert\bar{R}_4-\bar{R}_5\rvert = 17.25$
$\lvert\bar{R}_1-\bar{R}_5\rvert = 8.20$	$\lvert\bar{R}_2-\bar{R}_6\rvert = 35.95^*$	$\lvert\bar{R}_4-\bar{R}_6\rvert = 27.25^*$
$\lvert\bar{R}_1-\bar{R}_6\rvert = 18.20$	$\lvert\bar{R}_3-\bar{R}_4\rvert = 15.75$	$\lvert\bar{R}_5-\bar{R}_6\rvert = 10.00.$

Thus, 6 out of the 15 comparisons should be declared significant; these 6 significant comparisons are designated with asterisks.

Friedman's Test

Following Marascuilo and McSweeney (1967), consider n individuals or matched groups observed in a repeated measures design in which each subject or group is tested under

Table 1
Data for the Kruskal–Wallis-Type Example

Noise intensity level					
1	2	3	4	5	6
18 (10.5)	34 (40.5)	39 (49.5)	37 (44.5)	15 (8)	14 (7)
24 (20.5)	36 (43)	41 (51)	32 (34.5)	18 (10.5)	19 (12.5)
20 (14.5)	39 (49.5)	35 (42)	25 (22.5)	27 (25.5)	5 (1)
26 (24)	43 (54.5)	48 (58.5)	28 (28)	22 (17)	25 (22.5)
23 (18.5)	48 (58.5)	44 (56)	29 (30.5)	28 (28)	7 (2)
29 (30.5)	28 (28)	38 (47)	31 (33)	24 (20.5)	13 (5.5)
27 (25.5)	30 (32)	42 (52.5)	34 (40.5)	21 (16)	10 (3)
33 (37.5)	33 (37.5)	47 (57)	38 (47)	19 (12.5)	16 (9)
32 (34.5)	37 (44.5)	53 (60)	43 (54.5)	13 (5.5)	20 (14.5)
38 (47)	42 (52.5)	33 (37.5)	23 (18.5)	33 (37.5)	11 (4)

Note. Numbers in parentheses are the ranks associated with the original dependent variable scores.

k conditions. Let the original observations for each subject be replaced by ranks $(1, 2, \ldots, k)$ in accord with Friedman's two-way ANOVA procedure. Let $R_1, R_2, \ldots, R_k$ be the rank sums of each of the conditions, and let $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ be the respective average ranks.

From Marascuilo and McSweeney, when $E(\bar{R}_1) = E(\bar{R}_2) = \cdots = E(\bar{R}_k)$, it should be noted that if n is sufficiently large, the $\bar{R}_j$ will be approximately multivariate normal with

$$E(\bar{R}_j) = \frac{(k+1)}{2},$$

$$\text{var}\,(\bar{R}_j) = \sigma_j^2 = \frac{(k+1)(k-1)}{12n},$$

and

$$\text{cov}\,(\bar{R}_j, \bar{R}_{j'}) = \sigma_{jj'} = \frac{-(k+1)}{12n}$$

$$= \sigma_j^2\left[\frac{-1}{(k-1)}\right].$$

From Hartley's (1950) result follows the usual Tukey result that the probability is $1 - \alpha$ that all of the pairwise differences $[E(\bar{R}_j) - E(\bar{R}_{j'})]$ simultaneously satisfy

$$(\bar{R}_j - \bar{R}_{j'}) - \omega q_{\alpha,k,\infty} \leq [E(\bar{R}_j) - E(\bar{R}_{j'})] \leq (\bar{R}_j - \bar{R}_{j'}) + \omega q_{\alpha,k,\infty},$$

where $\omega = [k(k+1)/12n]^{\frac{1}{2}}$.

For data involving r sets of tied observations in which tied ranks are placed by their mean, it can be demonstrated that when $E(\bar{R}_1) = E(\bar{R}_2) = \cdots = E(\bar{R}_k)$, the $\bar{R}_j$ will be approximately multivariate normal if n is sufficiently large, with

$$E(\bar{R}_j) = \frac{(k+1)}{2},$$

$$\text{var}\,(\bar{R}_j) = \left(\frac{k-1}{nk}\right)\left[\frac{k(k+1)}{12} - \frac{\sum_{s=1}^{r}(t_s^3 - t_s)}{12n(k-1)}\right],$$

and

$$\text{cov}\,(\bar{R}_j, \bar{R}_{j'}) = \frac{-1}{(k-1)}\,\text{var}\,(\bar{R}_j).$$

Thus, for tied data, in accord with Hartley's result, it follows that the probability is $1 - \alpha$ that all possible pairwise comparisons $[E(\bar{R}_j) - E(\bar{R}_{j'})]$ simultaneously satisfy

$$(\bar{R}_j - \bar{R}_{j'}) - \omega^* q_{\alpha,k,\infty} \leq [E(\bar{R}_j) - E(\bar{R}_{j'})] \leq (\bar{R}_j - \bar{R}_{j'}) + \omega^* q_{\alpha,k,\infty},$$

where

$$\omega^* = \left[\frac{k(k+1)}{12n} - \frac{\sum_{s=1}^{r}(t_s^3 - t_s)}{12n^2(k-1)}\right]^{\frac{1}{2}}.$$

A Friedman-Type Example

Consider an example discussed in Hays (1973, p. 785). This hypothetical example involves a treatment variable with four levels. The treatments are applied to 11 groups of

Table 2
Data for the Friedman-Type Example

	Treatment level			
Group	1	2	3	4
1	1 (2)	4 (3)	8 (4)	0 (1)
2	2 (2)	3 (3)	13 (4)	1 (1)
3	10 (3)	0 (1)	11 (4)	3 (2)
4	12 (3)	11 (2)	13 (4)	10 (1)
5	1 (2)	3 (3)	10 (4)	0 (1)
6	10 (3)	3 (1)	11 (4)	9 (2)
7	4 (1)	12 (4)	10 (2)	11 (3)
8	10 (4)	4 (2)	5 (3)	3 (1)
9	10 (4)	4 (2)	9 (3)	3 (1)
10	14 (4)	4 (2)	7 (3)	2 (1)
11	3 (2)	2 (1)	4 (3)	13 (4)

Note. Numbers in parentheses are the ranks associated with the original dependent variable scores.

four matched subjects. The data for this example appear in Table 2.

An investigator wishes to test the six different hypotheses associated with pairwise comparisons of the form $[E(\bar{R}_j) - E(\bar{R}_{j'})] = 0$. Since there are no ties within any rows of the original data, ω should be calculated. For these data, $\omega = .39$. If the joint significance level of the six tests is to be controlled at $\alpha = .05$, one needs to obtain the value of $q_{.05,4,\infty}$; this value is 3.63. Those comparisons for which

$$|\bar{R}_j - \bar{R}_{j'}| > \omega q_{\alpha,k,\infty}$$

should be declared significant. In the present example, $\omega q_{\alpha,k,\infty} = 1.42$, and

$$|\bar{R}_1 - \bar{R}_2| = .55 \quad |\bar{R}_2 - \bar{R}_3| = 1.27$$
$$|\bar{R}_1 - \bar{R}_3| = .72 \quad |\bar{R}_2 - \bar{R}_4| = .54$$
$$|\bar{R}_1 - \bar{R}_4| = 1.09 \quad |\bar{R}_3 - \bar{R}_4| = 1.81^*.$$

Thus, only one out of the six comparisons should be declared significant; this comparison is designated with an asterisk.

Cochran's Test

Following Marascuilo and McSweeney (1967), consider n individuals observed in a repeated measures design in which each subject is tested under k conditions. In this case, the observations are dichotomous, for instance, *success* and *failure*. Let S_i be the number of successes for the ith subject. Let $T_1, T_2, \ldots, T_k$ be the sums for each condition, and let $\bar{T}_1, \bar{T}_2, \ldots, \bar{T}_k$ be the respective averages.

From Marascuilo and McSweeney, when $E(\bar{T}_1) = E(\bar{T}_2) = \cdots = E(\bar{T}_k)$, it should be noted that if n is sufficiently large, the $\bar{T}_j$ will be approximately multivariate normal with

$$E(\bar{T}_j) = \frac{\sum_{i=1}^{n} S_i}{nk},$$

$$\text{var}(\bar{T}_j) = \sigma_j^2 = \frac{1}{n^2} \sum_{i=1}^{n} \frac{S_i}{k}\left(1 - \frac{S_i}{k}\right),$$

and

$$\text{cov}(\bar{T}_j, \bar{T}_{j'}) = \sigma_{jj'} = \frac{-1}{(k-1)}\sigma_j^2.$$

From Hartley's (1950) result follows the usual Tukey result that the probability is $1 - \alpha$ that all of the pairwise differences $[E(\bar{T}_j) - E(\bar{T}_{j'})]$ simultaneously satisfy

$$(\bar{T}_j - \bar{T}_{j'}) - \omega q_{\alpha,k,\infty} \leq [E(\bar{T}_j) - E(\bar{T}_{j'})] \leq (\bar{T}_j - \bar{T}_{j'}) + \omega q_{\alpha,k,\infty},$$

where

$$\omega = \left\{\left[\frac{k}{(k-1)n^2}\right] \sum_{i=1}^{n} \frac{S_i}{k}\left(1 - \frac{S_i}{k}\right)\right\}^{\frac{1}{2}}.$$

Table 3
Data for the Cochran-Type Example

	Problem			
Subject	1	2	3	4
1	1	1	1	0
2	0	1	1	1
3	0	0	1	0
4	1	1	1	1
5	0	1	0	0
6	0	0	1	0
7	1	0	0	0
8	0	0	1	1
9	0	0	0	0
10	1	0	0	0
11	1	0	1	0
12	0	0	1	1
13	0	1	0	1
14	1	0	0	0
15	0	1	0	0
16	1	0	1	1
17	0	1	0	0
18	0	0	1	0
19	0	1	1	0
20	0	0	1	1

A Cochran-Type Example

Consider an example discussed in Hays (1973, p. 774). This hypothetical example involves 20 randomly selected subjects who were each given four problems in a random order. A 1 was recorded for a successful solution and a 0 for a failure. The data for this example appear in Table 3.

An investigator wishes to test the six different hypotheses associated with pairwise comparisons of the form $[E(\bar{T}_j) - E(\bar{T}_{j'})] = 0$. For these data, $\omega = .11$. If the joint significance level of the six tests is to be controlled at $\alpha = .05$, one needs to obtain the value of $q_{.05,4,\infty}$; this value is 3.63. Those comparisons for which

$$|\bar{T}_j - \bar{T}_{j'}| > \omega q_{\alpha,k,\infty}$$

should be declared significant. In the present example, $\omega q_{\alpha,k,\infty} = .40$, and

$$|\bar{T}_1 - \bar{T}_2| = .05 \quad |\bar{T}_2 - \bar{T}_3| = .20$$
$$|\bar{T}_1 - \bar{T}_3| = .25 \quad |\bar{T}_2 - \bar{T}_4| = .05$$
$$|\bar{T}_1 - \bar{T}_4| = .00 \quad |\bar{T}_3 - \bar{T}_4| = .25.$$

Thus, none of the six comparisons should be declared significant.

A Comment Concerning Sample Sizes

The analytic results presented in the present article obtain for large values of n. A further point that should be addressed concerns the question of what constitutes a large sample size for each of the cases considered here. Future empirical results would surely be helpful in answering such questions; however, there appears to be some information presently available that could serve as a useful guide. Siegel (1956) pointed out that the distribution of the usual Kruskal–Wallis test statistic can be closely approximated by an appropriate chi-square distribution when $n > 5$; Hays (1973) pointed out that the distribution of the usual Friedman test statistic can be closely approximated by an appropriate chi-square distribution when $n \geq 10$ and $k \geq 4$; and empirical results reported by Levy and Narula (1976) suggest that the distribution of the usual Cochran test statistic can be closely approximated by an appropriate chi-square distribution when $n \geq 18$. Although further work on these questions is needed, I suggest that the preceding statements provide useful information concerning the minimal values of large sample sizes for each of the cases considered in this article.

References

Cochran, W. G. The comparison of percentages in matched samples. *Biometrika*, 1950, *37*, 256–266.

Dunn, O. J. Multiple comparisons using rank sums. *Technometrics*, 1964, *6*, 241–252.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 1937, *32*, 675–701.

Hartley, H. O. The use of range in analysis of variance. *Biometrika*, 1950, *37*, 271–280.

Hays, W. L. *Statistics for the social sciences*. New York: Holt, Rinehart & Winston, 1973.

Kruskal, W. H., & Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 1952, *47*, 583–621.

Levy, K. J., & Narula, S. C. An empirical comparison of several methods for testing the equality of dependent proportions. *Communications in Statistics. Part B: Simulation and Computation*, 1976, *4*, 189–195.

Marascuilo, L. A., & McSweeney, M. Nonparametric post hoc comparisons for trend. *Psychological Bulletin*, 1967, *67*, 401–412.

Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw–Hill, 1956.

Received December 21, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 376–390

Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies

Sandra K. Mitchell
Department of Maternal and Child Nursing
University of Washington

Research in developmental and educational psychology has come to rely less on conventional psychometric tests and more on records of behavior made by human observers in natural and quasi-natural settings. Three coefficients that purport to reflect the quality of data collected in these observational studies are discussed: the interobserver agreement percentage, the reliability coefficient, and the generalizability coefficient. It is concluded that although high interobserver agreement is desirable in observational studies, high agreement alone is not sufficient to insure the quality of the data that are collected. Evidence of the reliability or generalizability of the data should also be reported. Further advantages of generalizability designs are discussed.

Almost everyone engaged in research recognizes the need for reliable measuring instruments. Reliability is a central topic particularly for courses and textbooks concerned with the behavioral sciences. In spite of varying theoretical derivations, its definition is remarkably uniform: A reliable instrument is one with small errors of measurement, one that shows stability, consistency, and dependability of scores for individuals on the trait, characteristic, or behavior being assessed.

Historically, the study of reliability has been linked to the study of individual differences and has been largely restricted to standardized tests of intelligence, achievement, and personality. These tests, however, are increasingly being replaced in developmental and educational psychology research by observations of subjects made in natural and quasi-natural settings. Although these observational studies vary widely in content and method, they all use human observers to record (and in some cases to summarize and abstract) the behavior of the subjects. Surprisingly, the reliability of these observational methods has not received the same attention as has the reliability of the more traditional methods (Johnson & Bolstad, 1973).

There are at least three different ways to think about the reliability of observational data. First, the researcher could focus on the extent to which two observers, working independently, agree on what behaviors are occurring. A coefficient that reflects the extent of this agreement has often been used to report reliability in observational studies. Second, the observational measure could be considered a special case of standardized psychological test, and the definitions of reliabil-

The preparation of this article was supported in part by Contract NO1-NU-14174 with the Division of Nursing, Bureau of Health Resources and Development, U.S. Public Health Service, Health Resources Administration, Department of Health, Education, and Welfare. The article is based in part on a dissertation submitted in partial fulfillment of the requirements for the PhD degree at the University of Washington.

I would like to thank Halbert B. Robinson and Kathryn E. Barnard for their support of the dissertation research, and Terence R. Mitchell and Nancy E. Jackson for their comments on an earlier draft of the article.

Requests for reprints should be sent to Sandra K. Mitchell, Department of Maternal and Child Nursing, Child Development and Mental Retardation Center, University of Washington, Seattle, Washington 98195.

ity that come from classic psychometric theory (e.g., test–retest and alternate forms) could be used. Finally, an observational measure could be thought to provide data that are under the influence of a number of different aspects of the observation situation (e.g., different observers or different occasions), including individual differences among subjects. This third viewpoint was developed in Cronbach's (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) theory of generalizability. The purpose of this article is to examine the appropriateness and correct interpretation of these three coefficients when they are used to reflect the quality and dependability of data gathered in observational studies.

Observer Agreement

To insure that data collected by human observers are objective, researchers typically obtain and report coefficients that demonstrate that two or more observers watching the same behavior at the same time will record the same data. These coefficients are offered as the reliability of the instrument being used.

Interobserver Agreement Percentage

The most common index of the quality of the data collected in observational studies is the interobserver agreement percentage. In its simplest form, this coefficient is just what its name implies: the percentage of time units during which the records of two observers are in agreement about the record of behavior.

Two comments can be made about the use of observer agreement percentages. First, the majority of studies report data only in this fashion. Consider, for example, the field of developmental psychology, in which about one third of recently published research articles use observational techniques. In Volume 47 (1976) of *Child Development*, 33 full-length articles reported observational measures. Of these studies, just under half (49%) reported only observer agreement figures as indicants of the quality of their data. Similarly, of 21 observational studies reported in Volume 12 (1976) of *Developmental Psychology*, 57% reported observer agreement percentages only.

Second, the amount of variability among the subjects in a study has very little impact, at least in theory, on the size of an interobserver agreement percentage. In actual practice, however, the degree of variability can make quite a difference: In very homogenous groups, observer agreement percentages are necessarily quite high because all scores given to all subjects are very close together. Thus, a measure that shows high agreement may, in some populations at least, have done a poor job of differentiating among subjects.

Other Problems With Observer Agreement

The interobserver agreement percentage has several important shortcomings. It is, first of all, insensitive to degrees of agreement, that is, it treats agreement as an all-or-none phenomenon, with no room for partial or incomplete agreement. In this sense, the percentage underestimates the actual extent of agreement between two observers. Second, some agreement between independent observers can be expected on the basis of chance alone. For many observational studies, especially those that use frequency counts of individual behaviors, the extent of this chance agreement is dependent on the rates at which the target behaviors occur. Behaviors with very high and very low frequencies can have extremely high chance levels of agreement (Johnson & Bolstad, 1973). In this sense, the percentage appears to overestimate the real agreement.

These difficulties in the use of observer agreement percentages are not unknown, and considerable effort has been expended to develop indices that overcome them. The alternative coefficients have been reviewed in detail elsewhere (Tinsley & Weiss, 1975). In general, they have been designed to overcome the mathematical shortcomings of the agreement percentage, such as chance levels of agreement.

A serious question remains, however, about the utility of these alternatives. In spite of their mathematical superiority, they deal with only one source of error (observer disagreement), and they deal with it without regard

to the magnitude of the individual differences. These alternatives may give a more accurate picture of the level of observer agreement in a study than does the simple percentage, but they do not otherwise describe the stability, consistency, and dependability of the data that have been collected.

Influences on Observer Agreement

Interobserver agreement has been experimentally studied as a phenomenon in its own right. In these experiments, observer accuracy, that is, agreement with a predetermined correct behavioral record, is the dependent variable.

Reid (1970) compared the accuracy of observers during overt and covert assessment of their reliability and found that they were significantly more accurate when they were aware that they were being checked. Romanczyk, Kent, Diament, and O'Leary (1973) found a similar drop from the overt to the covert assessment situation. They also found that observers recorded behavior differently depending on which of several researchers' records they thought their own records would be compared with. Taplin and Reid (1973) found that observer accuracy decreased between the end of a training period and the beginning of data collection and that it increased on days when "spot checks" were expected. Mash and McElwee (1974) reported that observers who had been trained to code "predictable" behavior (i.e., conversations with redundant information) showed a decline in accuracy when they later coded "unpredictable" behavior, whereas observers trained with unpredictable sequences showed no such decline in accuracy. Taken together, these studies imply that differences in experience, mental set, and training of observers can influence the accuracy with which behavioral records are made and scored.

Interobserver agreement has also been used as the dependent variable in studies that compare different methods of observation. McDowell (1973) found comparable observer agreement in time sampling and continuous recording of infant caretaking activities in an institution. Lytton (1973) found the interobserver agreement of ratings, home observations, and laboratory experiments to vary only slightly, but the amounts of time and effort necessary to achieve these levels were quite different. Mash and McElwee's (1974) rating system with only four categories was used more accurately than was an eight-category form. It appears, then, that there are differences in the interobserver agreement that can be expected from different methods of behavioral observation. These differences are reflected both in different levels of agreement for equal amounts of training and inequal levels of agreement for different amounts of training.

It is difficult, however, to put these agreement differences into perspective without knowledge of the overall variability among subjects in the studies. If between-subjects variability is low, then the reported differences in agreement due to observer instructions or observational methods may considerably influence the outcome of the study. If between-subjects variability is high, on the other hand, then these differences will probably have little influence. The studies that used agreement as a dependent variable have identified some problem areas in data collection, but they have not evaluated the relative importance of these problems.

Psychometric Theory of Reliability

The classical method of determining the reliability of a test is for the researcher to obtain two scores for a group of subjects on the test. These two scores may come from two separate scorings of the instrument, from administration of two parts or forms of the instrument to the subjects, or from two administrations of the same instrument to the subjects. The correlation between the two scores is the reliability of the instrument.

The central theoretical concept that underlies this psychometric view of reliability is that every test score is composed of two parts: a *true score,* which reflects the presence or extent of some trait, characteristic, or behavior, plus an *error score,* which is random and independent of the true score (Nunnally, 1967). The proportion of variance accounted

for by each of these parts is estimated from the correlation between the two scores obtained on the instrument.

The variance attributable to individual differences is usually given the same interpretation, regardless of how the two scores used to compute it were obtained. It reflects stable differences among individuals—the true score part of the data. The variance that is attributable to measurement error, however, is subject to varying interpretations, depending on how the two scores were obtained.

The error always includes, of course, the real error—those random fluctuations of the myriad factors that may affect the behavior being measured. These include such variables as the health or mental state of the subjects, the lighting or temperature in the testing room, and so forth. But the error also includes other sources of variation, depending on the method used to obtain the two scores. These other sources include differences within and between scorers, differences between different sections or forms of a test, and changes in subjects' behavior between two administrations of a test.

Thus, the three most common procedures for determining reliability involve (a) obtaining two separate scorings of the same instrument (intrascorer or interscorer reliability), (b) obtaining scores on two parts of the same instrument or on two very similar instruments (split-half or alternate-forms reliability), and (c) obtaining two scores from two separate administrations of the same instrument (test–retest reliability).

The researcher who wishes to use one of these classical methods of determining reliability in an observational study must somehow make his or her observations fit into the same general pattern as psychological tests. However, instead of one test with many items, all intended to measure the same trait, characteristic, or behavior, the observational researcher has a tool with a relatively small number of categories, with each category intended to measure a different trait, characteristic, or behavior. For each of these categories, which are generally mutually exclusive, data are usually collected during many distinct time units.

The most satisfactory way of making such data fit into the classic pattern seems to be to consider each mutually exclusive category (or type of behavior) a separate test with its own reliability. Each time unit is considered to be an item, since all time units are intended to measure the same trait, behavior, or characteristic. For example, a behavioral code might record a child's proximity to the teacher during each 10-sec unit of an observation. Each 10-sec unit would be an item in a test that measured proximity. If the measure consisted of a single summary score (such as in a rating), then there would be, in effect, no individual items at all, just one score.

This analogy between test reliability and the reliability of observational data can be extended to apply to each of the traditional ways of obtaining scores: intrascorer or interscorer reliability, split-half or alternate-forms reliability, and test–retest reliability.

Intrascorer or Interscorer Reliability

A clinical psychologist interested in self-directed aggression might listen twice to tape recordings of patients' responses to a projective test, each time counting the number of self-destructive statements. The correlation between the two counts for the group of patients would be the intrascorer reliability of the self-destructiveness score. The true score implied by this correlation would reflect real differences in self-destructiveness among the patients. The error would include not only the random error but also any inconsistencies in the psychologist's use of the self-destructiveness scale.

In actual practice, it is more likely that two psychologists would listen to the tape recordings. The correlation between their separate counts would be an interscorer reliability coefficient. The true score would again reflect real differences, but the error would reflect differences between the psychologists in their use of the scale, along with random error.

A similar situation exists, of course, when two or more observers record the behavior of subjects in other natural and quasi-natural settings. The correlation between the scores of two observers who kept track of how much

individual attention each child received from the teacher would be an interobserver reliability coefficient. This coefficient, once again, should not be confused with the observer agreement percentage.

Split-Half or Alternate-Forms Reliability

One way of determining the reliability of a standardized test is to compare scores on two subdivisions of the test (odd- and even-numbered items, frequently) or scores on two very similar versions of the test. In an observational study, the corresponding comparisons would be between subdivisions of one observation (e.g., odd- and even-numbered minutes during a tennis lesson), or between two very similar observations (first and second halves of a lesson, perhaps). This is an example of how time units can be considered analogous to test items.

Just as with interobserver reliability, the true score component of the variance in split-half or alternate-forms reliability reflects consistent individual differences among subjects. The error component, however, has a different interpretation. Along with random fluctuations in the behavior of the subjects, real differences in subject behavior between the two observed subdivisions are included as part of the error.

Test–Retest Reliability

Perhaps the most straightforward way to obtain two scores in a reliability study is to administer the same instrument at two different times. An observer might, for example, visit classrooms on different days to record the teacher's use of a particular instructional technique. As before, the true score is assumed to reflect some stable trait, characteristic, or behavior. In this case, the error includes not only random fluctuations of subject behavior but also whatever real changes in subject behavior have occurred between the two administrations of the test.

It is interesting to note that there is little difference between alternate-forms and test–retest reliability for observational measures. Since time units serve as items, observations made on different days can be considered either as alternate-forms or as test–retest conditions, depending on the situation.

Use of Reliability Coefficients

Three comments apply to all of these versions of the reliability coefficient. First, although the examples given are from hypothetical observational studies, real observational studies do not make use of all of the possible coefficients. Interobserver reliability or agreement is reported to the virtual exclusion of split-half and test–retest coefficients. Once again, developmental psychology can serve as an example. In Volume 47 (1976) of *Child Development*, 49% of the full-length research articles that used observational methods reported only observer agreement, and 39% more reported interobserver correlation coefficients. Only three of the studies (12%) reported a reliability coefficient that reflected the stability of subject behavior over time, that is, a split-half or test–retest reliability. Similarly, in Volume 12 (1976) of *Developmental Psychology*, 57% of the studies reported agreement only, 38% reported observer reliability coefficients, and only one study used a measure based on more than one sample of behavior per subject.

Second, although this discussion has emphasized the sources of error in these coefficients, the variance of true scores is as important in determining the size of the reliability coefficient as the variance of error scores is. Recall that the reliability of a test score can be expressed

$$\frac{\text{true score variance}}{\text{true score variance} + \text{error variance}}.$$

For a given level of error variance, then, an instrument will have a lower reliability when it is used on a homogenous group of subjects (low true score variance) than it will when it is used on a more heterogenous group (high true score variance). For instance, if the error variance is 10 and the true score variance is also 10, the reliability of the instrument is 10/20 or .50. But if the error variance is 10 and the true score variance is 40, the reliability of the instrument is 40/50 or .80. This is

in contrast to the observer agreement percentage, which is highest for homogenous groups of subjects.

Third, it should be repeated that reliability and observer agreement are not the same. It is possible, as illustrated by Tinsley and Weiss (1975), to have high interobserver agreement and a low reliability (correlation) coefficient, and vice versa. For instance, two observers might have perfect agreement about the color of shoes worn by children in a classroom, but if all the children wore red shoes, shoe color would not differentiate among the children. On the other hand, there might be a high correlation between two observers' records of the duration of a teacher's attention to a particular youngster, but if one observer's watch ran slower than the other's watch, they would probably never agree on the actual duration of the attention.

The differences between agreement and reliability are based on the way the two indices are defined. Reliability coefficients partition the variance of a set of scores into a true score (individual differences) and an error component. The error component may include random fluctuations in the behavior of subjects, inconsistencies in the use of the scale, differences among observers, and so forth. Interobserver agreement percentages, on the other hand, carry no information at all about individual differences among subjects and contain information about only one of the possible sources of error—differences among observers. In other words, a reliability coefficient reflects the relative magnitude of all error with respect to true score variability, whereas an agreement percentage reflects the absolute magnitude of just one kind of error.

All in all, there is no perfect reliability coefficient, nor is there one that is even generally best. Coefficients that use two scorings of the same instrument (interobserver and intraobserver reliability) confound random subject error with differences within and between scorers. Coefficients that use scores from subdivisions or alternative forms of the instruments (split-half and alternate-forms reliability) confound random subject error with differences between the subdivisions or forms. Finally, coefficients that use scores from the same instrument administered on two occasions (test–retest reliability) confound measurement errors with real changes in subject behavior that occur between the two administrations. The methods described cannot, then, separately estimate variance in test scores attributable to scorers, subtests (or forms), or occasions, nor can they consider these sources of error simultaneously. A more inclusive, multivariate theory is needed.

Generalizability Theory

Cronbach and his associates (Cronbach et al., 1972) have developed a theory that they call the theory of *generalizability*. (For a brief introduction to the theory, see J. P. Campbell, 1976, pp. 185–222.) Instead of assuming, as does classical test theory, that individual differences constitute the only lawful source of variation in test scores, generalizability theory assumes that there may be a number of sources of variation. These sources of variation other than individual differences are called *facets*. Different scorers, alternate test forms, or administration on different occasions are examples of facets that might be studied. A particular combination of facets makes up the *universe* to which test scores may be generalized.

A generalizability study (G study) is more reminiscent of a factorial study in experimental psychology than of a reliability study. In a G study, the researcher must collect data by systematically sampling conditions from each facet in the universe. For instance, two scorers might each score two alternate forms of a test given on different days to a group of subjects. Using an analysis of variance, it is then possible to independently estimate the contributions of each of the facets—scorers, forms, occasions, as well as subjects—to the overall variation in the set of test scores. Besides looking at the conventional F statistic to establish whether each facet makes a significant contribution to the scores, it is possible to compute what Cronbach calls *variance components*. These variance components reflect the size rather than the statistical significance of the contribution of each facet to the observed scores.

In examining the quality of observational data, though, not only are the absolute sizes of the variance components of interest but the relative sizes of the components are also important. The relative sizes, therefore, are the focus of this discussion.

Recall that a reliability coefficient reflects the partitioning of variance into true and error components and that the coefficient is the ratio of true score variance to obtained score variance. It represents, in other words, the proportion of the total variance that is accounted for by individual differences. In the same way, generalizability coefficient reflects the partitioning of variance into components that correspond to the facets sampled in the G study. The coefficient itself (an intraclass correlation) combines these components in a ratio that also represents the proportion of variance attributable to individual differences for a particular universe (set of conditions).

It should not be assumed, however, that a G study generates one coefficient that is appropriate for all applications of the instrument. On the contrary, one G study can generate several coefficients, each corresponding to a different universe of conditions. This fact points to an important distinction between the psychometric theory of reliability discussed earlier and the theory of generalizability. In psychometric theory, conditions of testing (or otherwise obtaining data) are assumed to influence only measurement error, not the true score on the instrument. In generalizability theory, on the other hand, the conditions of testing are assumed to influence the score itself. What Cronbach and his colleagues have shown is that while true scores all contain a common component, they also contain additional different components depending on the design; that is, it is not just the error variance that differs among the several reliability coefficients. This relationship can be illustrated by returning to the earlier example of interscorer reliability—two psychologists counting self-destructive statements from tape recordings. In generalizability terms, this is a one-facet study, that is, it samples observations from one facet (in this case, scorers) in addition to observations of different subjects. Analyzed as a G study, one can estimate variance components for subjects, for scorers, and for the interaction of subjects and scorers (which in this case is confounded with the residual error). The generalizability coefficient from this G study would reflect the dependability of a score for a subject generalized over scorers. In other words, it would indicate the proportion of variance accounted for by individual differences in subjects, above and beyond any effects accounted for by differences between scorers.

Suppose that a second facet—occasions—were added to this study, so that each scorer would count self-destructive statements for each tape two times. This study combines aspects of the interscorer and the intrascorer reliability studies. The analysis of this two-facet G study would yield variance components for subjects, for scorers, and for occasions (which in this case would be interpreted as intrascorer change). Further, variance components could be computed for each of the possible interactions of these facets: Subjects × Scorers, Subjects × Occasions, Scorers × Occasions, and Subjects × Scorers × Occasions (which in this case is confounded with residual error). The generalizability coefficient from this study would reflect the proportion of variance accounted for by individual differences in subjects apart from the effects between and within scorers.

Suppose further that this study were extended to three facets by having each psychologist use two different scoring methods for each tape recording he or she listened to (perhaps a count of self-destructive statements and a global rating of self-destructiveness). Observational methods would then be a facet in the universe of generalization.

Clearly each facet that is added to the study makes the information available from the analysis more complete. But there is a significant cost for the extra information provided by each facet: The number of observations required of each subject is multiplied by the number of conditions sampled in the facet. In the present example, the one-facet study would have two scores, the two-facet study would have four scores, and the three-

facet study would have eight scores for each subject.

Described below are some three-facet G studies that parallel the intraobserver–interobserver, split-half, and test–retest reliability studies that were discussed earlier. All of these G studies use the same three facets (observers, observational methods, and occasions) sampled for all subjects. The studies differ in their definitions of *occasions* of measurement and thus in their interpretations of the resulting coefficients. The relationships among the reliability studies and the proposed G studies are summarized in Table 1.

Table 1
Correspondence Between Reliability Studies and Generalizability Studies

Measurement occasions	Reliability study	Generalizability study
Separate scorings of the same behavior or instrument	Intrascorer or interscorer	Duplicate
Scores on two subdivisions of an instrument or behavior sample, or two very similar instruments or behavior samples	Split half or alternate forms	Session
Scores from separate administrations of the same instrument or separate samples of behavior	Test–retest	Developmental

Duplicate Generalizability

One study with this basic design might use audiotaped or videotaped recordings of behavior, which would be scored on more than one occasion by more than one observer, using more than one form of an observational instrument. In this G study (which I call *duplicate generalizability*), the occasions of observation actually consist of exactly the same behavior by the subjects. It is, then, an extension of the traditional intrascorer or interscorer reliability study. A reliability study uses two scores for each subject (usually from two different scorers) and confounds measurement error with differences within and between scorers. A duplicate G study, on the other hand, has many scores for each subject and separately estimates the contribution of differences within and among observers. Although *occasions* is a facet in this G study, variance attributable to occasions cannot be interpreted as within-subject change, since the same behavior occurs on each occasion of observation. In this study, occasions variance should be interpreted as a measure of within-observer stability. A duplicate G study would be appropriate for demonstrating the dependability of an instrument used in a study in which the stability of the behavior over time was not an issue.

Session Generalizability

A second G study with the basic design might be called *session generalizability*. It would use as measurement occasions two subdivisions of some behavioral sequence (i.e., first and second halves or odd and even minutes) and would be an extension of the traditional split-half reliability study. In a split-half reliability study, recall that errors of measurement are confounded with differences between the two halves of the test or observation. In a session G study, differences between subdivisions of the behavioral sequence are estimated by the variance component for occasions. A session G study would be used to estimate the dependability of scores reflecting traits and behaviors expected to be stable during the course of the behavioral sequence being observed, although perhaps no longer than that.

Developmental Generalizability

A third G study with the same basic design might be called *developmental generalizability*. It would use as measurement occasions two or more administrations of the same instrument, perhaps at different ages or developmental stages. It is, then, an extension of the

traditional test–retest reliability study, which confounds measurement error with true changes in behavior that have occurred between the two administrations of a test. In this developmental G study, these changes in behavior over time would be estimated by the occasions facet of the design. The developmental G study is best suited to measure of traits or characteristics believed to be relatively enduring.

Use of Generalizability Coefficients

The comments made about the reliability studies discussed earlier can be repeated for these G studies. First, observational studies very rarely report data in G-study terms. The method does appear occasionally in dissertations (see, e.g., Leler, 1971; Mitchell, 1977), and it surfaces now and then in educational psychology research (Medley & Mitzel, 1963; McGaw, Wardrop, & Bunda, 1972). But to return again to developmental psychology as an example, there were no studies published in 1976 in either *Child Development* or *Developmental Psychology* that reported generalizability coefficients. Second, regardless of the sizes of the variance components for the facets, it is necessary to have a relatively large variance component for subjects to obtain a large generalizability coefficient. All other things being equal, a sample of subjects with greater variability on the trait being measured will yield a higher generalizability coefficient than will a sample of subjects with lesser variability on the trait.

The three G-study designs outlined here (duplicate, session, and developmental) all sample three important facets that may influence scores in observational studies: observers, observational methods, and occasions of observation. They differ in the nature of the occasions that are sampled, and these occasions tell us something about the nature of the universe to which scores can be generalized.

One way to contrast these universes is to imagine that the three types of studies were conducted so that exactly the same subjects, observers, and observational methods were used for all three. When only the nature of the occasions facet is different, one can hypothesize certain relationships among the sizes of the coefficients derived from the three studies. When duplicate G studies are conducted, it is expected that variance due to occasions will be the smallest and hence that the generalizability coefficient will be the largest. Further, when developmental G studies are conducted, it is expected that variance due to occasions will be the greatest and that the generalizability coefficients will be the smallest. Finally, when session G studies are conducted it is expected that the variance due to occasions and the resulting generalizability coefficients will be intermediate.

Other Uses for Generalizability Theory

Although measures of interobserver agreement and reliability have important uses in observational research, it should be clear from the earlier discussion that it is the generalizability coefficient that is potentially the most useful source of information about the quality of such data. Generalizability theory, however, has many applications of interest for the developmental psychologist other than the computation of a coefficient. Some of these applications are discussed below.

Multitrait–Multimethod Matrix

D. T. Campbell and Fiske (1959) introduced the notion of determining the validity of psychological measurement instruments by using a *multitrait–multimethod matrix.* As the name suggests, this matrix consists of scores for an individual on several traits, each trait assessed by two or more different methods. Such a design is clearly an instance of a G study in which traits and methods are the two facets. The data analysis proposed by Campbell and Fiske uses a matrix of correlations, but other authors (i.e., Kavanagh, MacKinney, & Wolins, 1971) have used analyses of variance with the multitrait–multimethod design. In this form, such a matrix closely resembles a G study.

The multitrait–multimethod matrix was used by Wicker (1975) to examine the reliability of observational records generated

from transcriptions of conversations. In this study observers were treated as "methods," and behavior samples were treated as "traits." By applying Campbell and Fiske's criteria to the correlational matrix, Wicker concluded that his data showed both convergent and discriminant validity.

Attribution of Variance

Traditionally, psychological studies have sought either to demonstrate mean differences between groups of subjects or to show consistent individual differences among subjects. Another kind of study, far less common, tries to systematically apportion the variance in a set of research data among several independent variables. This approach has been popular in efforts to resolve the issue of whether individual differences (personality) or situational differences (environment) are the most important determinants of human behavior. In such studies, data are gathered on a group of individuals in several situations. The relative importance of individual differences and of situational differences is estimated by the use of the statistic known as omega-square. As Golding (1975) has ably demonstrated, this experimental design is more profitably viewed as a G study; generalizability coefficients answer questions about the relative importance of different facets (here, individuals and situations) more suitably than does omega-square.

The logic involved in this kind of study is not limited to the person–situation controversy, of course. A similar question might be asked about a study in which several raters rate the behavior of a number of subjects. As Norman and Goldberg (1966) pointed out, these data can be interpreted as reflecting the behavior of the ratees (subjects) or the behavior of the raters. Once again, this study appears to be a straightforward, one-facet generalizability study in which raters is the facet. Although Norman and Goldberg did not use analysis of variance to analyze their data either, it is clear that the conceptualization is similar to that presented earlier: There are several sources of meaningful variation in a set of data, and only a multifacet study can illuminate the relative contributions of different facets.

Observer Generalizability

The use of generalizability coefficients to estimate the contributions of facets other than individual differences to a set of test scores has a particular application to observational studies. Specifically, it allows a researcher to look at the proportion of variance in scores that is attributable to the consistent behavior of the observers.

On the surface, the function of the generalizability coefficient sounds much like the function of the interobserver agreement percentage, but in fact it is not. Recall that the agreement percentage did not take into account the extent of overall variability in a set of data, whereas a generalizability coefficient does. It should be noted that the G study necessary to compute this coefficient is exactly the same study as that used to compute the more conventional coefficient based on the behavior of the subjects. Nothing has changed except one's point of reference.

Mitchell (1977) computed both subject and observer generalizability coefficients in a study in which 67 observers made repeated observations of 10 mother–infant pairs during the first year of life. She found that the coefficients reflecting the variability accounted for by observers were, in this study at least, greater than the coefficients reflecting subject differences. Although this result is specific to this particular set of data, the study is an example of the usefulness of observer generalizability coefficients.

Single-Subject Studies

Studies of individual subjects have had few ways of reporting reliability in the traditional sense. However, it is possible to conduct G studies using only a single subject. For example, several observers, several occasions, or several methods of observation might be used with a single subject. In this case, the generalizability coefficient would reflect the generalizability of a score recorded by a single observer under the circumstances sampled in the

study; that is, it would be an observer generalizability coefficient.

Such single-subject studies may be appropriate even when many subjects are part of a research project. It is commonly assumed that the behavior of all subjects is recorded with equal accuracy, that is, measurement errors are approximately equal for all subjects. If this assumption is true, then the data for all subjects are presumably equally "good." On the other hand, if this assumption is incorrect, so that subject behavior is recorded with variable accuracy, then data for different subjects may have different meanings.

The possibility of systematic differences in measurement error among subjects has been explored for traditional psychological tests (Ghiselli, 1963). Berdie (1969), for example, found that intraindividual variability (i.e., measurement error) was a stable trait for some pencil-and-paper performance tests. Ghiselli (1960) examined the prediction of "predictability." He was able to predict the errors of measurement for two groups of subjects on a reaction time test. Reliability for the high group was .97, compared with .82 for the low group.

A similar result was found in a quite different study by Gorsuch, Henighan, and Barnard (1972). Their interest was the internal consistency of a children's pencil-and-paper test of locus of control. They found that the reliability of the scale differed significantly according to the reading ability of the children. The errors of measurement were quite small for the good readers, but were large for the poor ones.

Observational studies, however, rarely have enough subjects to permit analysis of this kind. An alternative is to make use of observer generalizability coefficients. Suppose that the data collected for each subject were considered to be a mini-G-study. If the basic G-study design outlined earlier were used, each mini-G-study would have two facets (methods and occasions) that would be sampled for each observer. From each of these mini-studies it would be possible to compute an observer generalizability coefficient. Using this design, Mitchell (1977) found that although there were differences in observer agreement for different subjects, subjects did not differ in their observer generalizability coefficients.

Summary

Three different coefficients that purport to reflect the quality of data gathered in observational studies have been discussed. The first and most commonly used of these was observer agreement. Coefficients of observer agreement are a source of important information about the quality of observational data: the objectivity of different observers using the same method to record the same behavior. Determination of interobserver agreement is a necessary part of the development and use of observational measures. Interobserver agreement is not, however, sufficient by itself.

The second coefficient discussed was the reliability coefficient, obtained by fitting observational data into the pattern used by developers of standardized psychological tests. There are really many different reliability coefficients, each defined by the way the scores are obtained for its computation. Reliability coefficients provide useful information about the stability and consistency of individual differences among subjects, but confound measurement error with other sources of variability.

The third coefficient was the generalizability coefficient, as defined by Cronbach et al.'s (1972) multivariate theory. In one sense, a generalizability coefficient supersedes a reliability coefficient, because it too provides information about the stability and consistency of individual differences among subjects. Its superiority to the reliability coefficient lies in its ability to account for variance from sources other than individual differences and measurement error. Besides giving information that can be reported as the generalizability coefficient, a G study also permits innovative ways of looking at results from observational studies. These ways include variations on the multitrait–multimethod matrix, the attribution of variance to independent variables, observer generalizability, and studies of single subjects. It is therefore especially unfortunate that G-study designs are not used more frequently in observational research.

Recommendations

Researchers doing observational studies are obliged to show that their measuring instruments are reliable—that they have small errors of measurement and that the scores of individuals show stability, consistency, and dependability for the trait, characteristic, or behavior being studied. The reason for this obligation is practical: If the measure is not reliable, it cannot be expected to show lawful relationships with other variables being studied. It is well-known that the reliability of a standardized test sets the limits of its validity (Nunnally, 1967). Similarly, the predictive usefulness of observational measures is limited by the stability and consistency of the scores obtained from the observational instruments.

Observer agreement coefficients alone, regardless of their mathematical sophistication, are inadequate to demonstrate this stability and consistency. What alternative or additional information ought to be reported in observational studies, and how should it be collected?

First, and most basically, the coefficients computed should be based on the same scores that are used in the substantive analysis of the study. If a composite score (such as a total of several categories or time units) is to be used for analysis, it is this composite—and not its component individual categories or time units—that should be examined for agreement, reliability, or generalizability. Of course, during the training of observers it is extremely helpful to compare the records of different observers on a trial-by-trial (or time unit by time unit) basis, but such a comparison does not suffice for reporting in a published research article. It is possible, albeit rare, to have acceptable agreement on a time unit basis and yet unacceptable levels of agreement on a total score. It is also possible, and much more common, for observers to be in only moderate agreement for small time units, but to show good agreement for a total score. In this case, an analysis of the trial-by-trial agreement would underestimate the agreement for the measures actually employed in the study.

Similarly, if several different scores are to be analyzed, coefficients should be computed for each of them. Good agreement or reliability on the frequency of particular behaviors, for example, does not insure good agreement or reliability on their duration. The data that are to be analyzed are the data that should be scrutinized for their stability and consistency.

Second, the coefficients should be computed from data that are part of the actual study being reported. The studies of observer agreement cited earlier (i.e., Reid, 1970; Romanczyk et al., 1973; Taplin & Reid, 1973) show clearly that the quality of data collected during a study may not be the same as the quality collected during reliability assessment or training. This difference can also be expected for data collected during a pilot study or during a previous study that used the same instrument.

This means that the researcher must plan to collect data that can be used to compute coefficients of agreement, reliability, or generalizability at the same time the rest of the data are gathered. Although this does entail some additional data collection, the addition need not generally be enormous (see, e.g., Rowley, 1976).

Third, interobserver agreement should always be obtained and reported. In most observational studies, observer disagreement is an important source of error, and it should be carefully and systematically monitored. This monitoring needs to be regular and unobtrusive for the most accurate results. Researchers also need to be alert to the possibility that interobserver agreement may differ for different subjects, therefore observations from many—preferably all—subjects should be used when determining the level of agreement.

Although the observer agreement percentage is the most widely used and most easily computed index of agreement, it may be desirable in some cases to substitute other indices, such as that suggested by Lawlis and Lu (1972). Whatever the exact form of the coefficient, however, both the researcher and the reader should remember that it reflects only one source of error and that it reports this error in absolute rather than in relative terms.

Fourth, a reliability or generalizability

coefficient that uses two or more measurement occasions should be presented for any score that is used to predict other behavior. The researcher is obliged to demonstrate that the individual differences among subjects are stable over different occasions as well as over different observers. This stability can be reported as a split-half, alternate-forms, or test–retest reliability coefficient, or as a session or developmental generalizability coefficient. The single exception to this rule is a study that focuses on some behavior not expected to show stability over time (e.g., first response to a new stimulus). In this case only, an interscorer reliability coefficient or a duplicate generalizability coefficient is appropriate.

It is impossible to overemphasize the importance of using two or more measurement occasions to compute coefficients of reliability or generalizability. The purpose of reporting such a coefficient is to demonstrate that the data being analyzed reflect stable, consistent, and dependable individual differences among subjects. If, however, a single measurement occasion has been used, then the coefficient can demonstrate only the competence and consistency of the observers. Since many, if not most, observational studies obtain repeated measures as part of the experimental design, it is seldom necessary to collect additional data. What is necessary is to analyze and report these observational measures in terms of their stability over time.

Fifth, a generalizability study is usually preferable to the computation of a reliability coefficient. First, a G study provides more useful information about sources of variability in a set of data than does a reliability coefficient. Leler (1971), for instance, used the variance components from a G-study analysis midway through her research to help refine the observation instrument and to retrain observers on some items. Second, the G-study design makes other kinds of analysis possible for most observational studies. These include, particularly, observer generalizability coefficients and coefficients for individual subjects.

Finally, the design of the generalizability study should correspond to the overall design of the research. The G study for most applications does not need to be complex. Two facets—observers and occasions—are usually sufficient. There are some studies, however, that require three-facet designs.

Studies that measure behavior before and after some intervention (experimental treatment) should include a third facet in the G study—before versus after the intervention. This is especially important if the intervention is likely to reduce the variability of the observed behavior. For example, suppose a study were undertaken to reduce the amount of aggressive behavior exhibited by school children on the playground. At the start of the study, the children would be quite variable in their playground aggression. If the intervention were successful, however, the variability after intervention would be quite low (all kids showing low aggression). A measure that would be quite reliable (in the classical sense) in differentiating among the children before the intervention might be inadequate afterwards. The G-study design allows the researcher to evaluate this possibility.

In the same way, studies that compare two groups of subjects should include group membership as a third facet in their G study; that is, it is necessary to demonstrate that the scores for both groups show approximately equal stability and consistency over different occasions and different observers.

Conclusions

These recommendations have an important empirical implication: Studies that follow them will report coefficients that are lower, perhaps substantially lower, than the coefficients reported by studies that do not follow them. The procedures suggested here are stringently conservative, and the coefficients that they yield should be considered lower limits of the true dependability of the observational data that are collected. Reviewers and readers, who are used to seeing reports of observer agreement in the .80s and .90s, will have to change their expectations for reliability and generalizability coefficients, which will often be in the .50s and .60s. In fact, these new coefficients are not low; rather the old ones were inappropriately high. Observer agreement percentages, interobserver reliabil-

ity, and reliability or agreement determined during pretests or previous studies are all spuriously high estimates of the quality of the data that are collected. Although we may have to revise our standards downward concerning the size of reported coefficients, we will be revising our standards upward concerning the ways in which the data for the coefficients are gathered and analyzed.

There is a methodological implication of these recommendations as well. Most observational studies currently being published are pretty straightforward in experimental design. Their sophistication is usually based on the nature of the observations themselves: the complexity of the behavioral record, the length of time included in the record, or the specific nature of the situation or setting in which the data are gathered. Future studies that follow the recommendations given here will be far more complex in design than is now typically the case.

One final question concerns whether it is really worth the additional time and money necessary to determine and report on the quality of the data in the ways suggested here. If including a G-study design in an observational study has no benefits except the computation of a generalizability coefficient, the answer is probably no. In fact, though, including a G study usually provides a great deal of substantive information to the researcher. Are there differences in the behavior reported by different observers? Do subjects act differently on different occasions? Is the variance of the data different before and after an experimental treatment? Are all groups in this study represented by data of equal quality? All these important questions can be addressed by including a G study in the overall research plan.

Even more importantly, though, the use of generalizability designs focuses the attention of the researcher (and the reader) on both the individual differences among subjects and on the influence of other (usually environmental) factors on behavior. As Cronbach (1957) eloquently pointed out, psychologists have historically tended to focus on one or the other of these two aspects: individual differences (using correlational methods) or group differences (using experimental methods). Fittingly, it is now Cronbach's theory of generalizability that makes it possible to combine these two viewpoints. And observational research is particularly well suited to the task of looking at individual differences in behavior in the context of systematic environment variation.

References

Berdie, R. F. Consistency and generalizability of intraindividual variability. *Journal of Applied Psychology,* 1969, *53,* 35–41.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81–105.

Campbell, J. P. Psychometric theory. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally, 1976.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist,* 1957, *12,* 671–684.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Ghiselli, E. E. The prediction of predictability. *Educational and Psychological Measurement,* 1960, *20,* 3–8.

Ghiselli, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology,* 1963, *47,* 81–86.

Golding, S. L. Flies in the ointment: Methodological problems in the analysis of the percentage of variance due to persons and situations. *Psychological Bulletin,* 1975, *82,* 278–281.

Gorsuch, R. L., Henighan, R. P., & Barnard, C. Locus of control: An example of dangers in using children's scales with children. *Child Development,* 1972, *43,* 579–590.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observations: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts and practice.* Champaign, Ill.: Research Press, 1973.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin,* 1971, *75,* 34–39.

Lawlis, G. F., & Lu, E. Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin,* 1972, *78,* 17–20.

Leler, H. O. Mother-child interaction and language performance in young disadvantaged Negro children (Doctoral dissertation, Stanford University, 1970). *Dissertation Abstracts International,* 1971, *31,* 4971B. (University Microfilms No. 71-2793)

Lytton, H. Three approaches to the study of parent-child interaction: Ethological, interview, and experimental. *Journal of Child Psychology and Psychiatry*, 1973, *14*, 1–17.

Mash, E. J., & McElwee, J. D. Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. *Child Development*, 1974, *45*, 367–377.

McDowell, E. E. Comparison of time-sampling and continuous recording techniques for observing developmental changes in caretaker and infant behaviors. *Journal of Genetic Psychology*, 1973, *123*, 99–105.

McGaw, B., Wardrop, J. L., & Bunda, M. A. Classroom observational schemes: Where are the errors? *American Educational Research Journal*, 1972, *9*, 13–27.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Mitchell, S. K. The reliability, generalizability and interobserver agreement of data collected in observational studies (Doctoral dissertation, University of Washington, 1976). *Dissertation Abstracts International*, 1977, *37*, 3583B. (University Microfilms No. 77-611)

Norman, W. T., & Goldberg, L. R. Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology*, 1966, *4*, 681–691.

Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.

Reid, J. B. Reliability assessment of observation data: A possible methodological problem. *Child Development*, 1970, *41*, 1143–1150.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. *Journal of Applied Behavior Analysis*, 1973, *6*, 175–184.

Rowley, G. L. The reliability of observational measures. *American Educational Research Journal*, 1976, *13*, 51–59.

Taplin, P. S., & Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. *Child Development*, 1973, *44*, 547–554.

Tinsley, H. E. A., & Weiss, D. J. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 1975, *22*, 358–376.

Wicker, A. W. An application of the multitrait-multimethod logic to the reliability of observational records. *Personality and Social Psychology Bulletin*, 1975, *1*, 575–579.

Received December 21, 1977 ▪

Psychological Bulletin
1979, Vol. 86, No. 2, 391–403

The Dichoptic Viewing Paradigm: Do the Eyes Have It?

Gerald M. Long
Stanford University

The two-part logic underlying the use of the dichoptic viewing procedure as a "psychoanatomical" tool by which to infer the retinal or cortical locus of critical processes for the perception of some visual phenomenon is reviewed and analyzed. Serious logical weaknesses are identified in both sides of the dichoptic argument: the inference of dominant retinal processes from unsuccessful dichoptic viewing or dominant cortical processes from successful dichoptic viewing. Specific examples from the visual literature are used to demonstrate the potential confounding of each of the variables suggested in this critique. The dichoptic viewing procedure tends to be employed in too uncritical a manner, and the usual interpretations of studies that have used this procedure as the sole technique by which to infer the general locus of a given visual phenomenon may be seriously in error.

Interest in localizing the physiological structures or levels within the visual system that are most directly involved in various perceptual phenomena has been directing research and theory since some of the earliest experimental work in vision (cf. Helmholtz, 1909/1962). However, this interest has taken on a renewed vitality in recent years with regard to the currently dominant information-processing conceptualization of perceptual functioning (e.g., Haber, 1969, 1974; Lindsay & Norman, 1972). In this now prevalent view, various stages and processes, usually in some form of hierarchical arrangement, by which stimulus information is transformed and transmitted in the visual system are hypothesized. Whether such stages purely represent intervening variables and thereby serve simply heuristic and shorthand functions for conceptualizing a cluster of logical processes or whether they actually refer to potential structures or groups of structures in the physiological system is frequently unspecified. Hence, the ultimate utility of these hypothetical processing stages in terms of a realistic conception of the physical system modeled is unclear. Julesz (1971) has recently provided an extremely lucid and detailed exposition of the value of a program of research aimed specifically at localizing within the visual system the physiological level(s) of processing involved in the perception of all manner of phenomena—from vernier acuity to eidetic memory to visual illusions. Julesz was particularly concerned with the potential roles random-dot stereograms and related techniques can play as psychoanatomical procedures by which the retinal or cortical locus of various visual percepts can be inferred from nonphysiological manipulations. However, a much older and in some sense related technique that is still very popular among researchers for ostensibly assessing the peripheral versus central locus of a visual percept is that of dichoptic viewing.

The author wishes to express his appreciation to B. Farrand, C. Krumhansl, N. Hersh, R. Schumer, and two anonymous reviewers for their helpful comments on earlier drafts of this article.

Requests for reprints should be sent to Gerald M. Long, who is now at the Department of Psychology, Villanova University, Villanova, Pennsylvania 19085.

Dichoptic Viewing Paradigm

Since at least the time of Helmholtz (1909/1962) it has generally been believed that the ability or inability of the visual system to combine information presented separately to

the two eyes can reflect directly the locus in the visual system of the processes involved. This dichoptic viewing procedure, which is sometimes also referred to as binocular or stereoscopic viewing, typically consists of the independent stimulation of the two eyes, either simultaneously or successively, by two aspects of the stimulus array that define some perceptual phenomenon.[1] Two rather well-known examples should clarify the procedures involved as well as the usual conclusions reached from this experimental manipulation. For a case of simultaneous dichoptic viewing, consider the interesting and sometimes controversial phenomenon known as *binocular yellow*. Under appropriate conditions, if a green stimulus is presented to one eye and an identical red stimulus is presented to the other eyes, the observer will frequently report a sensation of yellow. This is basically the same subjective sensation that occurs when both colored stimuli are presented to the same eye. The implications of this demonstration were clearly appreciated by early theorists in color vision. For Hecht (1928), this simple result exposed a critical weakness of the Hering (1964) and Ladd-Franklin (cited in Boring, 1942) theories, both of which required special retinal receptors for yellow light. On the other hand, the Young-Helmholtz three-receptor theory (cf. Boring, 1942) was sufficient to account for the phenomenon with the simple assumption that under the experimental conditions, "the red sensitive fibers and the green sensitive fibers are both active and the brain synthesizes yellow" (Hecht, 1928, p. 238). The seriousness of this attack for the other color theories was readily apparent to their proponents, and heated rebuttals to the dichoptic demonstration, stressing alleged procedural and stimulus weaknesses, were raised (Murray, 1930, 1939). Based on subsequent, more controlled experiments (e.g., F. H. Thomas, Dimmick, & Luria, 1961) and, of course, with the wisdom of hindsight bolstered by 40 intervening years of research, (e.g., P. K. Brown & Wald, 1964), it is not difficult to side now both with Hecht's argument for the authenticity of binocular yellow and the underlying three-pigment basis to normal color vision.

For an example of successive dichoptic viewing, there is the extensive literature on dichoptic visual masking (e.g., Turvey, 1973). In the usual dichoptic masking experiment, the target or test flash is presented to one eye, and the mask is presented, following a variable interstimulus interval, to the other eye. Perceptual performance (e.g., probability of detection or percentage of letters correctly reported) is then compared under the same temporal conditions with that obtained when both parts of the stimulus array, target and mask, are presented to the same eye (monocular viewing) or to both eyes (binocular viewing).[2] If there is very little difference between the dichoptic and monocular (or binocular) masking conditions, *postretinal* processes are usually inferred to underlie the masking effect. Hence, the masking of letters with complex patterns has been hypothesized to involve cortical processes, because of the general comparability of monocular and dichoptic demonstrations (cf. Breitmeyer & Ganz, 1976). On the other hand, if it is not possible to mask a target presented to one eye with a mask presented to the contralateral eye, *retinal* processes are inferred to dominate in any masking effects with the same stimuli obtainable under normal viewing conditions.

---

[1] A slight variation on this usual dichoptic approach, which is discussed at a later point in the article, should perhaps be mentioned here. Basically, it involves the presentation of the complete visual phenomenon to one eye (e.g., flickering light for critical flicker frequency determination) while some other stimulus is presented to the other eye (e.g., variable luminance background). If perceptual performance with the phenomenon is affected by the latter stimulus to the contralateral eye, central or cortical interactions are assumed to be involved (cf. J. L. Brown, 1966b).

[2] In some of the older literature, the term *binocular* was used in the same sense that *dichoptic* is used in the present context (e.g., Ammons & Weitz, 1951; Kahneman, Norman, & Kubovy, 1967). Correspondingly, phrases like *binocular interaction* or *binocular fusion* frequently referred to the visual system's processing of separate (i.e., dichoptic) stimuli presented to the two retinas. This usage of the word *binocular* appears subject to unnecessary confusion. Therefore, in the present article, *binocular* viewing is reserved for those conditions in which a single optic array is presented simultaneously to the two eyes, that is, normal viewing conditions.

Thus, homogeneous field masking (by a bright, blank field) has been relegated to intraretinal factors (cf. Breitmeyer & Ganz, 1976).

From these two examples, the reasons for the popularity of the dichoptic viewing paradigm are probably apparent. Theoretically, the two possible outcomes of such a procedure permit distinct conclusions: (a) If the visual phenomenon in question has remained essentially unchanged (although perhaps somewhat weakened in the dichoptic viewing condition), it is assumed that central (i.e., cortical) processes largely underlie the phenomenon, since only in postretinal centers is the separate input from the two eyes known to be combined; (b) on the other hand, if dichoptic presentation of a phenomenon destroys the usual percept, it is generally argued that the critical processes that determine the usual visual effect must be retinal or at least precortical. Peripheral processes are inferred, because the combination of the two half stimuli must apparently occur prior to the level of binocular innervation for adequate perception of the phenomenon in question. Needless to say, the latter conclusion that a given visual phenomenon is retinal does not mean that central stages are unnecessary; the perception of any event requires central structures. (In the extreme case, regardless of how perfect the functioning of the two eyes, if the cortical and subcortical structures are inoperative, one is blind to all visual phenomena.) Rather, the dichoptic viewing procedure has been employed to determine whether the central processing of the separate retinal signals from the half stimuli presented to different eyes is sufficient or whether peripheral structures must process certain aspects of the stimulus event in combination prior to the involvement of the higher centers. In other words, if a given visual event is not able to be perceived dichoptically, a nonlinear system has been inferred in which the result of the processing of A (input to left eye) + B (input to right eye) is not equivalent to the processing of A + B, the combined input to either eye alone.

Although the dichoptic viewing paradigm has had a considerable history, it was employed extensively by the Gestalt psychologists as the basis for an interocular criterion (Pastore, 1971) with which to demonstrate that more was involved in the perception of many phenomena than met the eye. Wertheimer (1912), for example, demonstrated that apparent motion was easily obtained even if the sequential stimuli were presented to separate eyes, thereby excluding interactions between adjacent retinal regions as a necessary basis for the phenomenon. Similarly, Köhler (1940; Köhler & Wallach, 1944) used the dichoptic paradigm in his work with figural aftereffects to argue that "the effect is located in the brain, where processes which are due to stimulation of corresponding parts of the two retinae occur in a common area" (1940, p. 86). However, these early references should not give the impression that the dichoptic procedure is in any way a paradigm of the past or that it is restricted to a very limited category of visual phenomena. The popularity of this psychoanatomical procedure by which one is theoretically able to infer the locus of critical processes for a large number of visual effects has remained quite strong up to the present day. Employing the two-part logic outlined in the previous paragraph, other researchers have used the dichoptic viewing paradigm to investigate the locus and nature of afterimages (e.g., Hansen, 1954), flicker discrimination (e.g., G. J. Thomas, 1955), apparent motion (e.g., Shipley, Kenney, & King, 1945), temporal summation (e.g., Kahneman, Norman, & Kubovy, 1967), Ganzfeld effects (e.g., Hochberg, Triebel, & Seaman, 1951), masking and metacontrast (e.g., Alpern, 1953; Turvey, 1973), geometric illusions (e.g., Papert, 1961; Schiller & Wiener, 1962), McCollough and related contingent aftereffects (e.g., Skowbo, Timney, Gentry, & Morant, 1975), short-term visual storage (e.g., Meyer, Lawson, & Cohen, 1975), frequency- or size-specific visual channels (e.g., Harter, Towle, & Musso, 1976), and many other visual phenomena. However, it should be noted that this technique's popularity has continued in spite of specific criticisms that have been raised against various aspects of the argument on which the dichoptic procedure is based. The present article attempts to outline these generally over-

looked weaknesses in the dichoptic procedure that serve to question subsequent conclusions concerning the locus of the critical processes underlying certain phenomena.

As a final comment before presenting these proposed weaknesses in the experimental procedure, it is of interest to note that previous critics of the dichoptic viewing paradigm appear to have focused on difficulties with one or the other of the two conclusions described above, while allowing the other to remain. For example, as is treated in more detail below, Julesz (1971, p. 6) tended to accept as probably "well-founded" the inference of critical cortical processes if the dichoptic viewing of some phenomenon is unimpaired, while rejecting the inference of dominant retinal processes if dichoptic viewing fails. Kolers (1972), on the contrary, argued that the dichoptic procedure only permits one to make the negative case:

> Failure to obtain a percept dichoptically proves only that most of the processing occurs early in the visual channel; but the occurrence of a percept dichoptically never proves the reverse, that all of its processing is normally cortical. (p. 182)

In other words, these two researchers will only accept the opposite conclusions from a dichoptic viewing experiment. In light of these differences of opinion concerning the legitimacy of either side of the dichoptic argument, it is the purpose of the present article to outline the rather strong case that can be made against drawing either of the usual conclusions from the dichoptic viewing experiment. Furthermore, to illustrate the points raised in this article, research examples from the current vision literature that appear potentially subject to some of the logical flaws to be described are employed.

The Negative Result: Does the Inability to Obtain a Percept Dichoptically Indicate Critical Retinal Processes?

Binocular (Retinal) Rivalry

As mentioned above, Julesz (1971) criticized the logic of inferring exclusive retinal processes as the basis of a given visual phenomenon if dichoptic viewing fails (i.e., if the normal percept cannot be obtained dichoptically). The reason for this involves the unknown role of binocular rivalry under dichoptic viewing conditions. It is well-known that if quite different information is presented to the two eyes, rather than simple fusion of the dichoptic images, the input from one eye may be partially or entirely suppressed while that of the other eye dominates (cf., Hochberg, 1972). Since it is very difficult to predict under which conditions fusion as opposed to suppression of separate retinal inputs will occur, binocular rivalry remains a serious confounding variable in the dichoptic viewing paradigm. Any change observed under dichoptic conditions may result from *interocular* suppression rather than from an assumed absence of critical *intraocular* processing presumably necessary for the usual percept.

Several indications of the potential involvement of binocular rivalry under conditions of dichoptic viewing can be found in the extensive masking literature. For example, Schiller and Smith (1968) reported greater masking effects in a metacontrast paradigm with dichoptic viewing than with monocular viewing. They interpreted this rather unexpected result in terms of binocular rivalry effects between the "test" eye and "mask" eye in addition to classic metacontrast suppression effects. Typically, influences from interocular rivalry are not considered in such work, but in this particular case, the additional hypothetical component of binocular rivalry was deemed necessary because the dichoptic suppression effect was larger than that of monocular suppression. Nevertheless, this does serve to emphasize the unknown contribution of binocular rivalry in those studies that report any effect of dichoptic viewing. Moreover, further evidence for ocular suppression during dichoptic viewing has been demonstrated convincingly in a recent study of masking effects by Monohan and Steronko (1977). In an attempt to control for binocular rivalry effects resulting from chronic ocular dominance of one eye over the other within subjects, these investigators preselected subjects on the criterion of equal masking effects across the two eyes before de-

termining luminance effects in a dichoptic pattern-masking experiment. Only 9 of 24 subjects tested met this criterion, but with these subjects results were obtained that were at odds with those previously reported in the masking literature. On the basis of these results, Monohan and Steronko argued for the reinvestigation of all reported dichoptic masking effects in which binocular rivalry, especially as reflected in ocular dominance, has not been controlled. It is proposed that this same argument should be given serious consideration with regard to other visual phenomena for which dichoptic viewing in any way alters performance.

In the context of the binocular rivalry argument, Julesz (1971) described the value of the random-dot stereograms that contain little monocular form information and thereby appear to eliminate or at least reduce binocular rivalry effects. To emphasize this point, Julesz criticized those studies investigating geometric illusions that inferred the dominance of retinal processes in these illusions because of the reduced illusory effect under dichoptic viewing conditions (e.g., Springbett, 1961). Following Day's (1961) argument that such findings may have resulted from binocular rivalry effects, Schiller and Wiener (1962) found that when brief tachistoscopic flashes of the geometric illusions were employed to reduce rivalry effects in the dichoptic situation, the magnitude of the illusory effects approached that obtained under binocular presentation. Using random-dot stereograms, Papert (1961) and Hochberg (cited in Julesz, 1971) reported these same illusions to be basically identical under classical (i.e., binocular) and cyclopean (i.e., stereogram) stimulation. Such results serve to confirm Julesz's argument that the earlier investigations were confounded by unknown effects of binocular rivalry under the dichoptic viewing conditions. However, it should be mentioned that even with the random-dot stereograms, the absence of successful stereoscopic viewing of some phenomenon only allows one "to conjecture its peripheral origin until it is proven to be central" (Julesz, 1971, p. 4). This caution is necessary because, although unimpaired dichoptic viewing may indeed reflect basic superposition of the retinal signals at central levels, the possibility of even partial binocular rivalry cannot be completely eliminated. The need for this caution is further supported by a recent study by Blake (1977) that demonstrated temporary suppression of an above-threshold target presented to one eye by a target at its own threshold contrast in the other eye. Hence, the extreme sensitivity of the mechanism that underlies ocular suppression appears to strengthen the argument against assuming an absence of binocular rivalry under any dichoptic viewing conditions.

Monocularly Innervated Cortical Effects

A second possible criticism of inferring retinal processes from a failure of dichoptic viewing is of a somewhat more hypothetical nature, but is nevertheless consistent with current physiological evidence. Hubel and Wiesel (1962, 1968) reported that the majority of cortical cells from which they recorded with microelectrodes were binocularly driven cells. However, a good proportion of their sample (up to 30%) appeared to be innervated only by one or the other eye. This, in turn, raises the possibility of some visual effects that may involve these monocularly driven cortical cells alone. Such visual effects would not be observable under conditions of dichoptic viewing, but would nevertheless be critically dependent upon postretinal processes. In support of this contention, it has recently been argued that color-contingent motion or orientation aftereffects may well fall into this category (Favreau & Corballis, 1976; Skowbo et al., 1975). In a typical demonstration of the color-contingent motion aftereffect (e.g., Mayhew & Anstis, 1972), a red clockwise-rotating spiral is alternated with a green counterclockwise-rotating spiral during the adaptation phase. In the subsequent test phase, if a stationary red spiral is viewed, it will appear to rotate counterclockwise; a stationary green spiral will appear to rotate clockwise. However, if during the adaptation phase the color is presented to the left eye and the rotating spiral to the right eye, no aftereffect is observed when the spiral

is subsequently viewed by the right eye (Murch, 1972). Hence, even though there is considerable evidence that these aftereffects are nonretinal (cf. Skowbo et al., 1975), it does not appear possible to innervate the appropriate cortical centers with the dichoptic presentation of the stimulus components during the adaptation phase. Whether future research will support these results with the dichoptic viewing of certain color-related aftereffects is not critical to the argument proposed here. The possibility of visual effects dependent even only in part on monocularly innervated cortical cells serves to weaken inferences of critical retinal processes that are based purely on the failure of dichoptic viewing.

The Positive Result: Does the Finding of No Difference Between Normal and Dichoptic Viewing Indicate the Relative Dominance of Central Processes?

Just as the previous two arguments have been raised against inferring dominant retinal processes from a failure of dichoptic viewing, it is also possible to find fault with the other side of the dichoptic paradigm by which cortical processes are inferred to play the critical role in the perception of some phenomenon if dichoptic viewing leaves a percept relatively unaffected. In this section various arguments, with examples from empirical work, are developed that seriously question this other half of the dichoptic paradigm.

Phenomenal Overlap of the Monocular Visual Fields

This first criticism can perhaps best be explained by reference to the controversy that existed for several years concerning the central versus peripheral locus of simple afterimages. As early as Newton, it was known that if an intense light source is viewed for a few seconds with one eye that is then closed and the nonstimulated eye is used for viewing, a negative afterimage of the source can sometimes be seen (e.g., Day, 1958; Terwilliger, 1963; Walls, 1953). This apparent *interocular transfer* of the afterimage was cited by some as evidence for central involvement in such afterimages (cf. Day, 1958). However, the bulk of both empirical and theoretical work favors the notion that the retina is the locus of classic afterimages (J. L. Brown, 1966a; Craik, 1940). If it is assumed that afterimages are due entirely to retinal fatigue, the problem arises as to the basis of the *phenomenal* transfer of such retinally based effects. In a footnote to his well-known study that investigated adaptation or aftereffects to curved lines, Gibson (1933) suggested the potential involvement of binocular rivalry in the apparent transfer of classic afterimages such that the stimulated but closed eye may dominate the open eye under some conditions. Sumner and Watts (1936) developed the argument more explicitly and found empirical support for this proposal by investigating differences in the apparent transfer of afterimages under various stimulus and background conditions. Day (1958) extended the argument even further and generalized its significance to all dichoptic viewing situations that involve aftereffects by stressing the possible confounding effect of the phenomenal overlap of the visual fields from the two eyes. In other words, Day argued that it is questionable to infer central processes from "successful" dichoptic viewing because it is not generally possible for the observer to separate truly centrally transferred effects from continued input from the closed but still signaling eye because of the overlap of the visual fields from the two eyes. This overlap of the two monocular fields has been proposed as the basis of the apparent transfer of negative afterimages, since the retinal pattern outlasts the physical stimulus whether the stimulated eye is open or closed following termination of the stimulus. Delabarre (1888) expressed strongly this same weakness of the dichoptic procedure 12 years before the turn of the century:

> A serious difficulty in settling the question [of the locus of afterimages] lies in the well-known impossibility of separating the visual fields of the two eyes. Whether one eye or both are open, whether they are focused on the same point or are held parallel, or squinted, or even jammed into all sorts of relative positions by fingers inserted into their sockets, the field of each will appear to coincide with the field of the corresponding portion of the retina of the other. (p. 326)

Much later, Terwilliger (1963) stressed the implications of these findings with afterimages for the study of other visual phenomena for which dichoptic viewing has proved successful. He argued that if a phenomenon with such a clearly retinal basis as afterimages exhibits introcular transfer, it does not appear theoretically sound to exclude categorically the involvement of retinal factors in any visual effects simply on the basis of successful dichoptic viewing. Terwilliger was specifically concerned with the possibility of an important retinal contribution to figural aftereffects, a notion rejected previously by Köhler and Wallach (1944) on the basis of their findings with dichoptic viewing described above. His methodological cautions, however, appear generally valid for all studies that employ the dichoptic procedure. Nonetheless, they are perhaps given further credibility by the subsequent emergence of alternate theories of figural aftereffects that stress the role of dominant retinal processes in such perceptual distortions (e.g., Deutsch, 1964; Ganz, 1966a, 1966b). Hence, as with afterimages, the use of the dichoptic viewing procedure may have retarded the development of an adequate theory of figural aftereffects by the uncritical inference of a retinally independent basis to the phenomenon solely on the basis of successful dichoptic viewing.

The potential confounding of the phenomenal overlap of the visual fields may seem rather obvious in retrospect, but it should be remembered that the question of the locus of afterimages was debated for years because of such apparent interocular transfer of the phenomenon. Furthermore, although it may appear implausible that this same problem could arise in current research, a much more subtle form of this criticism is less difficult to appreciate.

Hierarchical Cortical Processing of Retinal Signals

This less obvious—and therefore perhaps more dangerous—form of criticism against the inference of dominant cortical processes from successful dichoptic viewing was alluded to by Day (1958), but has been more recently and explicitly raised by Julesz (1971) and Sakitt (1976). It is a criticism based on the notion of successive stages within the visual system (i.e., levels of information processing) and rests on two logical premises. First, it must be remembered that one's ability to perceive a single unified (and three-dimensional) world from two separate retinas with overlapping visual fields results to a large degree from the eventual combinations of the two inputs at higher brain centers. This cyclopean conceptualization appears consistent both with current neurophysiological (e.g., Barlow, Blakemore, & Pettigrew, 1967; Pettigrew, 1972) and with psychophysical (e.g., Blake & Fox, 1973; Ono, Angus, & Gregor, 1977) evidence. Second, as mentioned earlier, it should be noted that one's awareness of some visual stimulus, no matter how completely dependent on retinal processes, results from the involvement of postretinal centers in the visual system. In the context of postulating the physiological basis for iconic memory, Sakitt (1976) made the distinction between the locus of the perception of some visual effect and the physical locus of the effect itself. Postretinal processes are necessarily involved in the former, yet a researcher frequently attempts to infer the latter from various experimental procedures such as the dichoptic viewing paradigm. To clarify further this distinction between the locus of a visual effect and the locus of the perception of that effect, consider the classic demonstration by Craik (1940) involving afterimages. He was able to show that although afterimages could be established without subjective awareness by temporarily pressure blinding the stimulated eye, higher centers were necessary for the awareness of the afterimage. Even retinal fatigue requires postretinal centers for subjective appreciation.

These two points simply reflect the basic fact that one's normal two-eyed perception of the world typically results from the eventual involvement of binocularly innervated cortical centers that in some way combine the inputs from the two retinas. Now, with regard to the dichoptic viewing paradigm, suppose a particular visual manipulation results in some change (unspecified) in the retinal signals in just one eye. These signals

will be transmitted to the higher centers in the visual system, where they may then interact with whatever signals are arriving from corresponding retinal regions of the other eye. Electrophysiological recording at the single-cell level in the striate cortex of the cat has demonstrated the existence of *binocular interaction fields* of excitatory and inhibitory regions that describe the activity of a simple cortical cell as a function of the relative stimulation of both eyes (Bishop, Henry, & Smith, 1971). The point is that dichoptic viewing of some visual phenomenon may appear to reflect critical central processes even though retinal factors may dominate and largely define the phenomenon. This results from the basic fact that one's awareness of the visual world necessarily involves postretinal structures, and in these successive postretinal stages of information processing, monocular inputs are eventually combined. Hence, an experimental procedure that can tap directly only postretinal processing (i.e., observers' verbal reports) may overestimate the cortical contributions to some visual phenomena because of the interaction between incoming retinal signals. This is a serious weakness of the dichoptic viewing procedure.

To illustrate the potential confounding of this relatively late combined processing of retinal inputs, it may be best to focus on those studies that have employed a slight variation of the usual dichoptic procedure (see Footnote 1). In these studies, some complete visual pattern is presented to one eye, and the effect of introducing some other stimulus to the other eye is determined. If this latter stimulus has some effect on the perception of the stimulus to the first eye, purely central processes are inferred to be involved. For example, Jacewitz and Lehmann (1972) presented a typical Sperling (1960) partial-report task to an observer's left eye while varying the input to the right eye from that of a blank field to a train of either blank or grid flashes. For the partial-report task, nine letters in a 3 × 3 array were presented for 50 msec, followed at some brief delay by a variable-pitch tone that directed the observer to report the top, middle, or bottom row of letters only. The average number of letters available to the observer (i.e., the percentage of the three letters correctly reported per trial times the total number [nine] presented per trial) was then estimated for each time delay condition and as a function of the three visual conditions for the contralateral eye. The decrease in recall performance with increasing complexity of the contraocular signals was interpreted in terms of reduced central processing capacity available; and, hence, the cortical locus of the iconic memory theoretically assessed by the partial-report procedure was concluded. Similar logic underlies the use of an analogous procedure in the study of critical flicker frequency (CFF). Variations in the values of CFF for a given intermittent stimulus presented to one eye when the input to the other eye is varied "are presumed to reflect central interaction processes" (J. L. Brown, 1966b, p. 259).

In some cases, the conclusion of dominant cortical processes from these studies may well be correct. However, the above brief quote by Brown reveals a potential flaw. What is central, the phenomenon itself, or just the interaction with the contraocular signals? Did Jacewitz and Lehmann in the study described previously demonstrate the cortical locus of the persistence required on the iconic memory task or the cortical locus of the interference from the other retinal signals? These uncertainties can be stated more formally within an information-processing framework (cf. Haber, 1974). Given a hierarchical arrangement of processing stages (both retinal and cortical) within the visual system, the dichoptic procedure cannot clearly distinguish between visual effects whose dominant processes are cortical and those whose processes are only partly cortical or even those for which retinal processes are most critical. Some examples from the current visual literature should help to clarify this interpretive difficulty.

Example 1: Spatial frequency channels in the visual system. Consider the recent study by Harter et al. (1976) in which different checkerboard patterns of various sizes were presented dichoptically to the two eyes. Visual evoked potentials and reaction times to a checkerboard pattern that was flashed to the left eye were determined as a function of the checkerboard size in that pattern and as

the size of a constant checkerboard pattern viewed by the right eye was varied. Both response measures indicated reduced sensitivity to a pattern presented to the left eye when that pattern was most similar to the pattern viewed by the right eye. The authors concluded that the results were consistent with the notion of binocularly innervated, size-specific cortical neurons, which were selectively adapted by the pattern presented to the right eye.

It is not the purpose of the present discussion to dispute the specific conclusions of the Harter et al. study. As they pointed out, there is considerable converging evidence from other studies for the same conclusion; so, in fact, there may be good reason to accept their proposal. Rather, the importance of the Harter et al. study in the present context is to demonstrate the plausibility of an alternate explanation of dichoptic results based upon the particular weakness of the dichoptic procedure that concerns the ultimate combination of retinal signals prior to an observer's response. In this regard, consider the predicted results if there were only retinal size-specific cells—perhaps the retinal ganglion cells with their concentric receptive fields of varying size. A fatiguing of one such set of cells in the right eye could result in increased noise or inhibitory signals being transmitted to the higher centers, thereby affecting the ultimately perceived input from the left eye. Hence, given the phenomenal overlap of the monocular visual fields and the eventual convergence of ocular signals at cortical levels prior to awareness, it would be possible to explain the same pattern of results in terms of size-specific retinal processes. Unique processing by cortical centers need not be inferred. Moreover, consider the likely pattern of results if the above study had employed simple homogeneous fields of varying luminance levels also presented dichoptically. Results similar to those of Harter et al. that reflected a depression in performance with increasing test (left eye) and background (right eye) similarity would most likely be expected, but it is highly doubtful that such results would be interpreted in terms of the existence of binocularly innervated, luminance-specific cortical neurons. Before proposing such a construct, a researcher would first wish to rule out in some way the known retinal effects of varying luminance.

Several other studies that probed the nature of the hypothesized spatial frequency (or size-specific) channels in the visual system also employed dichoptic techniques as experimental means by which to infer the locus of such effects. Blakemore and Campbell (1969) and Blakemore and Sutton (1969) demonstrated the interocular transfer of adaptation effects that were spatial frequency specific, theoretically reflecting the adaptation of a narrowly tuned, binocularly innervated cortical channel. More recently, Blake and Levinson (1977) demonstrated interocular facilitation of grating detection (i.e., lowered threshold contrast for a monocular, striped target by the presentation of a subthreshold grating to the other eye). This effect also was spatial frequency specific: Optimal facilitation was found when the gratings to the two eyes were most similar. Although these studies may indeed share some of the weaknesses of the dichoptic procedure detailed previously, there are other critical factors that render their dichoptic procedure considerably less suspect. On the other hand, as mentioned in the context of the Harter et al. (1976) study, there is significant converging evidence both from other psychophysical studies and from electrophysiological investigations that supports the cortical locus of neurons in the visual system that are selectively sensitive to the spatial frequency of a retinal stimulus (cf. Blake & Levinson, 1977). However, even more important is the fact that in the three studies just cited, an additional empirical result was also stressed that substantiates the postretinal interpretation of the locus of such effects. Specifically, the interocular effects described in these studies were also found to demonstrate *orientational selectivity;* that is, only if the contraocular stimulus was of roughly the same orientation (tilt) as the target stimulus was the interocular viewing condition comparable to that of monocular viewing. The importance of this additional demonstration for the conclusion of a central locus to the effects rests in the known types of receptive fields of retinal cells (cf. Kuffler

& Nicholls, 1976). The last stage of processing within the retina, that of the ganglion cells, exhibits concentric receptive fields of various sizes. Therefore, purely retinal processing of a stimulus may demonstrate some spatial frequency analysis either by the differential excitation of ganglion cells by stimuli of varying widths or by the combined activity of a large number of ganglion cells with randomly distributed receptive fields (Kelly, 1975). Later, binocularly innervated cortical cells could simply reflect the processing properties of this earlier stage in the system. However, orientational selectivity of these same perceptual effects would not be expected until the involvement of higher centers that combine the output from several such concentric ganglion cells at specific orientations (cf. Hubel & Wiesel, 1962). Hence, the repeated finding in the literature of successful dichoptic (i.e., interocular) viewing is not in itself conclusive evidence for a characteristically or qualitatively different nature to the processing at earlier (retinal) levels in the system prior to the cortical conjunction of monocular signals; but the complementary finding of orientational selectivity renders this same conclusion much more tenable. It should be stressed that the dichoptic procedure alone is not the convincing demonstration.

These examples from the visual literature that deals with hypothesized spatial frequency channels were intended to exemplify an inherent weakness of the dichoptic procedure that results from the hierarchical arrangement of the visual system. With the important exception of orientational selectivity, it does not appear readily apparent how, through the use of the dichoptic procedure alone, a researcher could distinguish between binocular centers that serve largely as relay stations for retinally processed signals and binocular centers that actively process and transform the retinal signals. It is realized that in some respects this argument may strike the reader as somewhat hollow, considering the other available evidence that supports the central locus of these spatial frequency channels (cf. Blake & Levinson, 1977; Breitmeyer & Ganz, 1976). But it should be remembered that the focus of the present argument is not on the specific conclusions of these studies but on their methodology. Julesz (1971) similarly discussed this procedural weakness of the dichoptic procedure, and Kolers's (1972) argument against inferring a necessary cortical locus from the unimpaired dichoptic viewing of some phenomenon rested on the same points (Kolers, Note 1). The importance of this logical error can be demonstrated perhaps more persuasively by reference to an area of research that, like afterimages and aftereffects, may have been retarded by the uncritical and unquestioned use of the dichoptic procedure.

Example 2: Locus of iconic memory. Iconic memory, or short-term visual storage, refers to the persistence exhibited by the visual system following the brief presentation of a target (cf. Neisser, 1967; Sakitt, 1976). It has been argued in the past that the locus of iconic memory must be postretinal because of the demonstrated dichoptic masking of the icon (cf. Dick, 1974; Jacewitz & Lehmann, 1972). However, there is increasing evidence from recent research that the bulk of what has been referred to in the literature as iconic memory involves retinal persistence effects (Long, 1978; Sakitt, 1976; Sakitt & Long, 1978). In her review of the iconic literature, Sakitt (1976) suggested that the peripheral locus of the icon can be reconciled with the results of dichoptic masking experiments if it is assumed that the prolonged signals from the retina are transmitted to the higher levels where binocular interaction occurs. Hence, although the masking itself may take place at a central site of binocular combination, the icons themselves may be due entirely to prolonged photoreceptor activity. As mentioned previously, since the perception of any event occurs not in the eye itself but at higher levels, it is necessarily difficult using the dichoptic procedure to tease out the visual phenomena (e.g., afterimages) that have their predominant impact on the retina (which in turn sends its output to binocularly innervated cortical centers) from the more purely central phenomena that are not contained in the monocular signals alone.

A similar argument can also be raised against the conclusion by Haber and Standing (1969) of the central locus of iconic memory

from results of dichoptic presentation of stimuli in a nonmasking paradigm. In their study, a circular target of variable duration was repeatedly presented at various rates until the subject reported the target to be on continuously. The presentation of the target to alternate eyes on successive flashes (i.e., dichoptic viewing) had no effect on the chosen temporal interval for continuity of perception, as compared with monocular viewing conditions. This lack of difference between monocular and dichoptic viewing of the target was interpreted as evidence for the central locus of the persistence effect. However, in light of the present argument it is claimed that such results actually demonstrate only that the input from the two eyes is combined prior to the perceptual decision about the subjective duration. The results do not indicate whether the persistence occurred prior to or at this point of combination; they show only that the perception of the event occurred after the confluence of the retinal signals.

Value of "correlograms" for localizing processes. The criticisms raised in this section against the inference of a cortical basis to phenomena that can be perceived dichoptically evince serious limitations to the practical value of the typical dichoptic demonstration as an unambiguous psychoanatomical tool. However, it should be noted that these criticisms do not appear to be equally damaging to the related nonphysiological procedures represented by random-dot stereograms. Julesz (1971) painstakingly outlined the value of random-dot stereograms, anaglyphs, and other forms of so-called correlograms in inferring cortical processes for a large range of visual phenomena. The usual criticisms raised against drawing this cortical inference from dichoptic demonstrations are much less applicable to this rather intriguing class of stimuli. Each monocular pattern by itself is drastically inadequate for perception of the given phenomenon and appears to each eye as composed of random dots. It is only in the particular horizontal disparities between portions of the two retinal inputs that the complete visual phenomenon is defined. In Julesz's (1971) words, "Random-dot stereograms do not contain, *even physically* the global information in the left and right retinal projections. It is only the *relation* between the left and right patterns that produces a pattern of the desired kind" (p. 7). Since this comparison between retinal inputs (i.e., disparity detection) can only occur after the locus of conjunction of monocular images, cortical processes must underlie the perception of these phenomena that are portrayed exclusively on the cyclopean retina of the mind's eye.[3]

Summary

The purpose of the present article has been to point out several potential weaknesses that underlie the use of the dichoptic viewing procedure to infer peripheral versus central processes in the perception of a given phenomenon. These problems with the dichoptic procedure were shown to limit the conclusions from either the successful (i.e., unimpaired) or the unsuccessful dichoptic viewing of some visual effect. It was argued that failure in dichoptic viewing, instead of necessarily reflecting the retinal locus of critical processes for the perception of a given phenomenon, could also result from either binocular rivalry effects or the sufficiency of monocularly driven cortical events for a certain percept. On the other hand, the successful dichoptic viewing of a visual phenomenon, instead of indicating the dominant cortical locus of the phenomenon, could result from the phenomenal overlap of the visual fields (such that the observer cannot distinguish actual transfer from continued output from the stimulated eye) and from the relatively

[3] An interesting recent example of the use of such cyclopean stimuli for localizing processes can be found in a study by Fox and Lehmkuhle (Note 2). Dynamic noise stereograms were used to present the brief array of letters in a Sperling (1960) partial-report task. Each monocular pattern alone appeared as randomly moving elements in the display; only in the disparity between certain identical portions in the two monocular inputs were the letters contained (i.e., seen in depth). No iconic memory was found for these purely postretinal letters. Not only is this result consistent with other recent findings that indicate a peripheral locus to iconic memory (e.g., Sakitt & Long, 1978) but it is also opposite to those results, described previously, that were obtained with traditional dichoptic procedures.

late stage of combination of the input from the two eyes. To illustrate each of these arguments, empirical examples were employed to demonstrate the plausibility of each alternate hypothesis to the classic dichoptic interpretation. It is believed that the criticisms raised in this article seriously question the previously uncritical and almost automatic use of the dichoptic viewing procedure for determining the level of the visual system that underlies performance on a given visual task.

Reference Notes

1. Kolers, P. A. Personal communication, summer 1977.
2. Fox, R., & Lehmkuhle, S. *Iconic memory in stereospace: Seeing without storing*. Paper presented at the meeting of the Psychonomic Society, Washington, D.C., November 1977.

References

Alpern, M. Metacontrast. *Journal of the Optical Society of America*, 1953, *43*, 648–657.

Ammons, C. H., & Weitz, J. Central and peripheral factors in the phi phenomenon. *Journal of Experimental Psychology*, 1951, *42*, 327–332.

Barlow, H. B., Blakemore, C., & Pettigrew, J. D. The neural mechanism of binocular depth discrimination. *Journal of Physiology*, 1967, *193*, 327–342.

Bishop, P. O., Henry, G. H., & Smith, C. J. Binocular interaction fields of single units in the cat striate cortex. *Journal of Physiology*, 1971, *216*, 39–68.

Blake, R. Threshold conditions for binocular rivalry. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 251–257.

Blake, R., & Fox, R. The psychophysical inquiry into binocular summation. *Perception & Psychophysics*, 1973, *14*, 161–185.

Blake, R., & Levinson, E. Spatial properties of binocular neurones in the human visual system. *Experimental Brain Research*, 1977, *27*, 221–232.

Blakemore, C., & Campbell, F. W. On the existence of neurones in the human visual system selectively sensitive to orientation and size of retinal image. *Journal of Physiology*, 1969, *203*, 237–260.

Blakemore, C., & Sutton, P. Size adaptation: A new aftereffect. *Science*, 1969, *166*, 245–247.

Boring, E. G. *Sensation and perception in the history of experimental psychology*. New York: Appleton-Century, 1942.

Breitmeyer, B. G., & Ganz, L. Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing *Psychological Review*, 1976, *83*, 1–36.

Brown, J. L. Afterimages. In C. J. Graham (Ed.), *Vision and visual perception*. New York: Wiley, 1966. (a)

Brown, J. L. Flicker and intermittent stimulation. In C. H. Graham (Ed.), *Vision and visual perception*. New York: Wiley, 1966. (b)

Brown, P. K., & Wald, G. Visual pigment in single rods and cones of the human retina. *Science*, 1964, *144*, 45–52.

Craik, K. J. K. Origin of visual afterimages. *Nature*, 1940, *145*, 512.

Day, R. H. On interocular transfer and the central origin of visual aftereffects. *American Journal of Psychology*, 1958, *71*, 784–789.

Day, R. H. On the stereoscopic observation of geometric illusions. *Perceptual and Motor Skills*, 1961, *13*, 247–258.

Delabarre, E. B. On the seat of optical after-images. *American Journal of Psychology*, 1888, *2*, 326–328.

Deutsch, J. A. Neurophysiological contrast phenomena and figural aftereffects. *Psychological Review*, 1964, *71*, 19–36.

Dick, A. O. Iconic memory and its relation to perceptual processing and other memory mechanisms. *Perceptions & Psychophysics*, 1974, *16*, 575–596.

Favreau, O. E., & Corballis, M. C. Negative aftereffects in visual perception. *Scientific American*, 1976, *235*(6), 42–48.

Ganz, L. Is the figural aftereffect an aftereffect? A review of its intensity, onset, decay, and transfer characteristics. *Psychological Bulletin*, 1966, *66*, 151–165. (a)

Ganz, L. The mechanism of figural aftereffect. *Psychological Review*, 1966, *73*, 128–150. (b)

Gibson, J. J. Adaptation, after-effect and contrast in the perception of curved lines. *Journal of Experimental Psychology*, 1933, *16*, 1–31.

Haber, R. N. *Information-processing approaches to visual perception*. New York: Holt, Rinehart & Winston, 1969.

Haber, R. N. Information processing. In E. C. Carterette & M. P. Friedman (Eds.), *Historical and philosophical roots of perception*. New York: Academic Press, 1974.

Haber, R. N., & Standing, L. G. Direct measures of short-term visual storage. *Quarterly Journal of Experimental Psychology*, 1969, *21*, 43–54.

Hansen, A. D. "After-image transfer test" in anomalous retinal correspondence. *Archives of Ophthalmology*, 1954, *32*, 369–374.

Harter, M. R., Towle, V. L., & Musso, M. F. Size specificity and interocular suppression: Monocular evoked potentials and reaction times. *Vision Research*, 1976, *16*, 1111–1117.

Hecht, S. On the binocular fusion of colors and its relation to theories of color vision. *Proceedings of the National Academy of Sciences*, 1928, *14*, 237–241.

Helmholtz, H. von. [*Physiological optics*] (J. P. C. Southall, Ed. and trans.). New York: Dover, 1962. (Originally published, 1909.)

Hering, E. *Outlines of a theory of the light sense*. Cambridge, Mass.: Harvard University Press, 1964.

Hochberg, J. Perception: II. Space and movement. In J. W. Kling & L. A. Riggs (Eds.), *Woodworth & Schlosberg's Experimental psychology*. New York: Holt, Rinehart & Winston, 1972.

Hochberg, J., Triebel, W., & Seaman, G. Color adaptation under conditions of homogeneous stimulation (Ganzfield). *Journal of Experimental Psychology*, 1951, *41*, 153–159.

Hubel, D. H., & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 1962, *160*, 106–154.

Hubel, D. H., & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 1968, *195*, 215–243.

Jacewitz, M. M., & Lehmann, D. Iconic memory, dichoptic interference and short-term consolidation. *Neuropsychologia*, 1972, *10*, 193–198.

Julesz, B. *Foundations of cyclopean perception*. Chicago: University of Chicago Press, 1971.

Kahneman, D., Norman, J., & Kubovy, M. Critical duration for the resolution of form: Centrally or peripherally determined? *Journal of Experimental Psychology*, 1967, *73*, 323–327.

Kelly, D. H. Spatial frequency selectivity in the retina. *Vision Research*, 1975, *15*, 665–672.

Köhler, W. *Dynamics in psychology*. New York: Liveright, 1940.

Köhler, W., & Wallach, H. Figural aftereffects: An investigation of visual processes. *Proceedings of the American Philosophical Society*, 1944, *88*, 269–357.

Kolers, P. A. *Aspects of motion perception*. New York: Pergamon Press, 1972.

Kuffler, S. W., & Nicholls, J. G. *From neuron to brain*. Sunderland, Mass.: Sinauer Associates, 1976.

Lindsay, P. H., & Norman, D. A. *Human information processing: An introduction to psychology*. New York: Academic Press, 1972.

Long, G. M. *Iconic memory: The effects of stimulus parameters on short-term visual storage*. Unpublished doctoral dissertation, Stanford University, 1978.

Mayhew, J. E. W., & Anstis, S. M. Movement aftereffects contingent on color, intensity, and pattern. *Perception & Psychophysics*, 1972, *12*, 77–85.

Meyer, G. E., Lawson, R., & Cohen, W. The effects of orientation-specific adaptation on the duration of short-term visual storage. *Vision Research*, 1975, *15*, 569–572.

Monohan, J. S., & Steronko, R. J. Stimulus luminance and dichoptic pattern masking. *Vision Research*, 1977, *17*, 385–390.

Murch, G. M. Binocular relationships in a size and color orientation specific aftereffect. *Journal of Experimental Psychology*, 1972, *93*, 30–34.

Murray, E. Color problems: The divergent outlook of physicist and psychologist. *American Journal of Psychology*, 1930, *42*, 117–127.

Murray, E. Binocular fusion and the locus of "yellow." *American Journal of Psychology*, 1939, *52*, 117–121.

Neisser, U. *Cognitive psychology*. Englewood Cliffs, N.J.: Prentice-Hall, 1967.

Ono, H., Angus, R., & Gregor, P. Binocular single vision achieved by fusion and suppression. *Perception & Psychophysics*, 1977, *21*, 513–521.

Papert, S. Centrally produced geometrical illusions. *Nature*, 1961, *191*, 733.

Pastore, N. *Selective history of theories of visual perception: 1650–1950*. New York: Oxford University Press, 1971.

Pettigrew, J. D. The neurophysiology of binocular vision. *Scientific American*, 1972, *227*(2), 84–95.

Sakitt, B. S. Iconic memory. *Psychological Review*, 1976, *83*, 257–276.

Sakitt, B. S., & Long, G. M. Relative rod and cone contributions in iconic storage. *Perception & Psychophysics*, 1978, *23*, 527–536.

Schiller, P. H., & Smith, M. C. Monoptic and dichoptic metacontrast. *Perception & Psychophysics*, 1968, *3*, 237–239.

Schiller, P. H., & Wiener, M. Binocular and stereoscopic viewing of geometric illusions. *Perceptual and Motor Skills*, 1962, *15*, 739–747.

Shipley, W. C., Kenney, F. A., & King, M. E. Beta apparent movement under binocular, monocular and interocular stimulation. *American Journal of Psychology*, 1945, *58*, 545–549.

Skowbo, D., Timney, B. N., Gentry, T. A., & Morant, R. B. McCollough effects: Experimental findings and theoretical accounts. *Psychological Bulletin*, 1975, *82*, 497–510.

Sperling, G. The information available in brief visual presentation. *Psychological Monographs*, 1960, *74* (11, Whole No. 498).

Springbett, B. M. Some stereoscopic phenomena and their implications. *British Journal of Psychology*, 1961, *52*, 105–109.

Sumner, F. C., & Watts, F. P. Rivalry between uniocular negative after-images and the vision of the other eye. *American Journal of Psychology*, 1936, *48*, 109–116.

Terwilliger, R. F. Evidence for a relationship between figural aftereffects and afterimages. *American Journal of Psychology*, 1963, *76*, 306–310.

Thomas, F. H., Dimmick, F. L., & Luria, S. M. A study of binocular color mixture. *Vision Research*, 1961, *1*, 108–120.

Thomas, G. J. A comparison of uniocular and binocular critical flicker frequencies: Simultaneous and alternate flashes. *American Journal of Psychology*, 1955, *68*, 37–53.

Turvey, M. T. On peripheral and central processes in vision: Inferences from an information-processing analysis. *Psychological Review*, 1973, *80*, 1–52.

Walls, G. L. Interocular transfer of after-images. *American Journal of Optometry*, 1953, *30*, 57–64.

Wertheimer, M. Experimental studien uber das sehen von bewegung. *Zeitshrift für Psychologie*, 1912, *61*, 161–265.

Received December 30, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 404–419

Categories for Classifying Language in Psychotherapy

Robert L. Russell
Department of Linguistics
University of North Carolina at Chapel Hill

William B. Stiles
University of North Carolina at Chapel Hill

A review of language analysis systems employed in psychotherapy research suggests a typology based on the combination of three category types with two coding strategies. The types are (a) content categories, (b) intersubjective categories, and (c) extralinguistic categories. They are defined by distinct sets of language features. The coding strategies are (a) the classical coding strategy, in which categories describe the text, and (b) the pragmatic coding strategy, in which categories describe the speaker. A review of research results suggests that the content, intersubjective, and extralinguistic features constitute distinct channels of communication and that (a) the content channel carries information pertaining to the speaker's psychodynamic process and personality structure, (b) the intersubjective channel carries information pertaining to the quality of the speaker's relationship with the other, and (c) the extralinguistic channel carries information pertaining to the speaker's transitory emotional state. System consistency criteria are suggested for use in conjunction with the typology to evaluate categories and category systems.

If the first stage in the scientific study of a phenomenon is naming and classifying, the study of verbal behavior in psychotherapy is mired in its first stage. Reviews of the psychotherapy content analysis literature (Auld & Murray, 1955; Kiesler, 1973; Marsden, 1965, 1971; Meltzoff & Kornreich, 1970) display an ungainly proliferation of categories and systems of categories to describe the verbal behavior of therapist and client. One reviewer (Kiesler, 1973) put it this way:

> Psychotherapy process research has to rank near the forefront of research disciplines characterized as chaotic, prolific, unconnected, and disjointed, with researchers unaware of much of the work that has preceded and the individual investigator tending to start anew completely ignorant of closely related previous work. (p. xvii)

The proliferation of categories and category systems and the disorganization of this area of research reflect the lack of consensus on what are the most significant aspects of verbal interaction. In the absence of a unified theoretical understanding, investigators seem dissatisfied with existing systems, and they continue to search for more illuminating ways to capture the richness of verbal behavior in psychotherapy (Goodman & Dooley, 1976; Kepecs, 1977; Kiesler, 1973; Labov & Fanshel, 1977; Strupp, 1957; Stiles, in press-b). The results of this continuing search—the numerous categories and classification systems—have created a need for an integrated descriptive framework, a framework that would facilitate the comparison and evaluation of alternative categorization schemes (Freedman, Leary, Ossorio, & Coffey, 1951–1952; Reusch & Bateson, 1949; Rice, 1965; Rice & Wagstaff, 1967). The present article addresses this need; that is, it describes an order we see among existing language analysis categories and suggests guidelines for selecting or creating classification systems appropriate for particular research problems.

Our framework is a typology of categories. We use three distinct sets of language features to identify three types of categories.

The authors thank John B. Carroll and Bruce C. Johnson for their comments on a draft of this article.

Requests for reprints should be sent to Robert L. Russell, Department of Linguistics, University of North Carolina, Chapel Hill, North Carolina 27514.

The types we propose subsume the vast majority of the language categories that have been used to study verbal interaction in psychotherapy. In brief, *content* categories, such as *mother* or *death anxiety,* concern denotative or connotative semantic content. *Intersubjective* categories, such as *self-disclosure* or *question,* concern syntactically implied and other relationships between the communicator and recipient. *Extralinguistic* categories, such as *pauses* or *laughing* concern vocal noises, tonal qualities, and temporal patterning of speech, defined independently of semantic content and syntactic structures.

Cutting across the three types are two distinct coding strategies, previously described by Berelson (1952) and by Marsden (1965, 1971). In the *classical* strategy, categories describe characteristics of the text (or some other record of the communication), whereas in the *pragmatic* strategy, categories describe characteristics of the communicator such as his or her internal state, intentions, socioeconomic class, and so on. For example, the category *mother* would be classical if it coded instances of maternal references in the text; the category *death anxiety* would be pragmatic if it coded utterances judged to reflect the communicator's conscious or unconscious concern about death. The classical strategy requires two operational steps from the raw text to inferences about psychological processes: The coder identifies instances of categories in the text, and the researcher later makes inferences based on category frequencies (or indices derived from category frequencies). The classical strategy thus makes explicit the process of inference about the communicator's characteristics. The pragmatic strategy uses only one step; coders make inferences about psychological processes (or other characteristics of the communicator) in the process of coding. These inferences may or may not be based on specified behaviors (e.g., behaviors that instantiate death anxiety may or may not be exhaustively catalogued), but in either case, the specific behaviors are not recorded. Thus the pragmatic strategy permits complex contextual judgments that may be impossible to specify completely (Labov & Fanshel, 1977; Russell, in press), but it obscures the relationships between behaviors and inferred characteristics of the communicator (Marsden, 1971).

The classical–pragmatic distinction has been used for distinguishing whole systems of categories or approaches to psychotherapy process research (Berelson, 1952; Marsden, 1965, 1971), but it is better applied to characterize individual categories: Single systems can (and do) contain both classical and pragmatic categories. Likewise, although the classical–pragmatic distinction has been used primarily for systems that employ content categories, it is useful for intersubjective and extralinguistic categories as well.

Previous efforts at organizing language research in psychotherapy have classified studies or systems by criteria other than category types. Auld and Murray (1955) distinguished methodological studies, descriptive studies of cases, and theoretically guided studies of therapy. Kiesler (1973) distinguished "systems of direct psychotherapy process analysis," which focus on therapist and/or patient behavior, from "systems of indirect psychotherapy analysis," which include indirect process analysis, therapist's conceptions, and patient preferences. Marsden (1965, 1971) partitioned studies into three models, the classical, the pragmatic, and the nonquantitative. As noted above, we have adopted the classical–pragmatic distinction to classify categories. Several other authors have made distinctions that parallel aspects of our typology. Mahl and Schulze (1964), Matarazzo and Wiens (1977), and Phillips, Matarazzo, Matarazzo, Saslow, and Kanfer (1961) have distinguished extralinguistic categories and category systems from all other category and category system types. Phillips et al. recognized two controlling frames of reference: one that "directs attention to the communication aspects of verbal behavior, that is, to some symbolic content of the words spoken, using content analysis to define and quantify its variables" and one that "focuses upon the quantitative temporal characteristics of interview interaction, utilizing measures such as number and duration of utterances, duration of silence, etc." (p. 260). Mahl and Schulze (1964) reviewed just those studies

concerned with the extralinguistic features of speech (e.g., pitch, pauses, rhythm, etc.). Similarly, several investigators (Dollard & Auld, 1959; Murray, 1956; Seeman, 1949; Snyder, 1945, 1963) have implicitly recognized a distinction between content categories and intersubjective categories by constructing systems of content categories for patient verbal behavior and separate systems of intersubjective categories (called *technique* categories by Dollard & Auld, 1959) for therapist verbal responses.

Our 3 × 2 typology is not an exhaustive classification of methods used to analyze verbal interaction in psychotherapy. The typology is not intended to cover ratings of verbal behavior, though ratings clearly cut across our typology. Thus, for example, ratings of client "experiencing" (Klein, Mathieu, Gendlin, & Kiesler, 1970; Rogers, 1958, 1959) or the "immediacy" of therapist or patient responses (Mehrabian, 1972) are not included. Also, the typology is not intended to cover systems that classify language units marked

Table 1
Content Category Systems

Murray (1956); patient content categories pragmatic strategy

Disturbance of free association
Agreement with therapist remarks [intersubjective]
Disagreement with therapist remarks [intersubjective]
Intellectual discussion
General anxiety
Sex
Sex anxiety
Sex frustration
Affection
Affection anxiety
Affection frustration
Dependence
Dependence anxiety
Dependence frustration
Independence and self-assertion
Independence anxiety
Independence frustration
Unspecified anxiety
Unspecified frustration

Stone, Dunphy, Smith, & Ogilvie (1966); Harvard Third Psychosociological Dictionary (partial list); classical strategy

Natural realm; psychological processes; emotions
- Arousal—states of emotional excitement
- Urge—drive states
- Affection—incidents of close . . . relationships
- Pleasure—states of gratification
- Distress—states of despair, guilt, shame, etc.
- Anger—forms of aggressive expression

Thought
- Sense—perceptions and awareness
- Think—cognitive processes
- If—conditional words
- Equal—words denoting similarity
- Not—words denoting negation
- Cause—words denoting a cause–effect relationship

Evaluation
- Good—synonyms for good
- Bad—synonyms for bad
- Ought—words indicating a moral imperative

Dollard & Auld (1959); patient's signs; pragmatic strategy

An internal response-produced stimulus; anxiety, apprehension, distress, tension, or fear
Unconscious anxiety or unconscious sense of guilt
Anxiety is perceptibly reduced
Reduction of unconscious anxiety or guilt
Confirmation
Dependence
Unconscious dependence
Patient is aware of and frightened by his dependent motive
Conscious dependent motive to which unconscious anxiety is attached
Dependent motive is unconscious; anxiety is conscious
Both dependent motive and anxiety are unconscious
Anxiety component is reduced
Unconscious anxiety attached to conscious dependence motive is reduced
Dependence motive is unconscious; conscious anxiety component is reduced
Unconscious anxiety evoked by unconscious dependence motive is reduced

White, Fichtenbaum, & Dollard (1966b); patient's evaluation of self categories; pragmatic strategy

Positive self-evaluation
Anxiety about self, symptom, dissatisfaction
Hostile feelings directed toward others; assertiveness
Hostile feelings directed toward self; self-blame
Nonsexual love and reduced anxiety about family
Anxiety evoked by family members
Sex motive, dating, marital relations
Anxiety evoked by sex motive
Involvement with academic and career motive; social mobility
Anxiety evoked by academic/career motive or social mobility

Table 1 (*continued*)

Kepecs (1977); focal conflict categories; pragmatic strategy	Laffal (1968); Cognitive Conceptual Dictionary (partial list); classical strategy
Positive human relations	Absurd
Hostility out	Agree 1 (sympathy)
Mastery	Agree 2 (agreement)
Assertion	Agree 3 (similarity)
Reactions (reactive motive)	All 1 (whole)
Hostility in	All 2 (much)
Masochism	All 3 (frequent)
Helplessness	Animal
Other defenses	Art
Defensive hostility out	Astronomy 1 (space)
Defensive positive human relations	Astronomy 2 (weather)
Defensive mastery	Back
Adaptive activity; relatively nonconflictual solutions	Bad
	Begin
	Big
Hall & Van de Castle (1966); dream contents (partial list); classical strategy	**Gottschalk & Gleser (1969); Anxiety Content Analysis Scale; pragmatic strategy**
Objects	Death anxiety
Architecture	Mutilation anxiety
Household	Separation anxiety
Food	Guilt anxiety
Implements	Shame anxiety
Travel	Diffuse or nonspecific anxiety
Streets	**Thibaut & Coules (1952); overt aggression categories; pragmatic strategy**
Regions	Direct aggression
Nature	Indirect aggression
Body parts	Affective neutrality
Clothing	Friendly statements
Communication	Self-augmentation
Money	Self-reduction
Miscellaneous	Self-neutral

Note. Discrepant category types are listed in brackets.

for some special syntactic feature such as case (Bieber, Patton, & Fuhriman, 1977; Patton, Fuhriman, & Bieber, 1977). The categories we discuss are mainly nominal variables, so indices of intensity or degree are based on frequencies. We think it best to concentrate on the category types that appear most productive and recur most frequently in psychotherapy process research rather than to attempt to include all possible categories. On the other hand, the typology we propose may also be useful for areas other than psychotherapy research (see Berelson, 1952; for an introduction to applications of language analysis categories to other fields in the social sciences, see Gerbner, Holsti, Krippendorf, Paisley, & Stone, 1969; Holsti, 1968; Pool, 1959).

The category systems shown in Tables 1, 2, and 3 were classified as content, intersubjective, or extralinguistic, respectively, on the basis of predominant category type in the system. Similarly, they were classified as pragmatic or classical on the basis of the predominant coding strategy that was employed to score speech units to the constituent categories.

Content Categories

Content categories describe the semantic content of words or word groups in the text.

Table 2
Intersubjective Category Systems

Stiles (in press-a); verbal response modes

Classical strategy
- Disclosure form
- Question form
- Edification form
- Acknowledgment form
- Advisement form
- Interpretation form
- Confirmation form
- Reflection form

Pragmatic strategy
- Disclosure intent
- Question intent
- Edification intent
- Acknowledgment intent
- Advisement intent
- Interpretation intent
- Confirmation intent
- Reflection intent

Murray (1956); therapist content; pragmatic strategy

Instructions
Labels
Strong approvals
Disapprovals
Demands
Directions
Mild probes
Mild approvals
Mm [classical strategy]
Not classifiable [residual]

Bales (1970); interaction process analysis; pragmatic strategy

Seems friendly [mixed]
Dramatizes [mixed]
Agrees
Gives suggestion
Gives opinion
Gives information
Asks for information
Asks for opinion
Asks for suggestion
Disagrees
Shows tension [extralinguistic]
Seems unfriendly [mixed]

Porter (1943); therapist checklist; pragmatic strategy

Defining the interview situation [mixed content and intersubjective]
Bringing out and developing the problem situation; leading
Developing client's insight and understanding; clarification, interpretation, and problem identification
Sponsoring client activity; fostering decision making

Snyder (1945); counselor categories; pragmatic strategy

Structuring
Forcing client to choose and develop topic
Directive questions
Nondirective leads and questions
Simple acceptance
Restatement of content or problem
Clarification or recognition of feeling
Interpretation
Approval and encouragement
Giving information or explanation
Proposing client activity
Disapproval and criticism

Lennard & Bernstein (1960); therapist informational specificity categories; pragmatic strategy

Passive encouragement
Active encouragement
Limits to subjective matter area
Limits to specific old proposition
Interpretation
Limits to specific answer
Excludes discussion
Introduces specific new proposition

Strupp (1957); type of therapeutic activity; pragmatic strategy

Facilitating communication
Exploratory operations
Clarification
Interpretive operations
Structuring [content]
Direct guidance
Activity not clearly related to the task of therapy [content]
Unclassifiable [residual]

Bandura, Lipsher, & Miller (1960); therapist activity; pragmatic strategy

Reflection
Labeling
Exploration
Approval
Ignoring
Topical transition [content]
Silence [classical extralinguistic]
Mislabeling

Goodman & Dooley (1976); listener response modes; pragmatic strategy

Question [classical strategy]
Advisement
Interpretation
Reflection
Disclosure
Silence [classical extralinguistic]

Note. Discrepant category types are listed in brackets.

Table 3
Extralinguistic Category Systems

Dibner (1956); speech characteristics; classical strategy Unfinished sentence Breaking in with a new thought, generally by breaking into another sentence [mixed] Interrupted sentence Repeating words or phrases Stuttering I don't know [content] Sighing Laughing Voice change Questioning the interviewer [intersubjective] Blocking	Mahl (1956); speech disturbance categories; classical strategy "Ah" Sentence correction Sentence incompletion Repetition Stutter Intruding incoherent sound Tongue slip Omission
Rice & Wagstaff (1967); voice quality; pragmatic strategy Emotional Focused Externalizing Limited	Eldred & Price (1958); voice categories; classical strategy Alterations of pitch: overhigh and overlow Alterations of volume: overloud and oversoft Alterations of rate: overfast and overslow Breakup
	Lasswell (1935); speech categories; pragmatic strategy Slower speech rate; increase of unconscious tension Faster speech rate; decrease of unconscious tension
Matarazzo, Wiens, Matarazzo, & Saslow (1968); interaction chronograph; classical strategy Mean speech duration Mean speech latency Percentage of interruption	Fairbanks & Pronovost (1939); pitch categories; classical strategy Pitch level Pitch range Extent of pitch shifts

Note. Discrepant category types are listed in brackets.

(Examples of content category systems are given in Table 1.) The categories describe manifest or latent content: Denotative meaning and connotative meaning constitute manifest content; referential–contextual, symbolic, or metaphorical meanings constitute latent content. For example, Lasswell (quoted in Kaplan, 1943) wrote, "'I love my husband more than anyone in the whole world' may be taken at its face value; or we may decide that the wife 'doth protest too much.' In the first instance we describe manifest content, and in the second, we interpret according to latent content" (p. 234). Manifest content is identified by the classical coding strategy, in which the coder makes inferences about the characteristics of the communicator.

Classical content categories describe the manifest content of the text; that is, either the dictionary meanings of the words or word groups that occur in the text or the connotative or constituent meanings normally ascribed to them, regardless of the context in which they occur. Thus classical content categories serve "as a conceptual grid to be laid upon a language sample in order to reveal the density of the various concepts in the sample" (Laffal, 1968, p. 280). Classical content categories vary in abstractness. The following is an example of a concrete category (also see Table 1):

Implements: Three subclasses of implements are scored. The first letter of the scoring symbol is I to which a second letter is attached to indicate the subclass.

Tools (Scoring symbol: IT). This subclass includes tools, and machinery parts. Objects that are used in vocational activities are generally included here, although some such as typewriter are scored in the communication class. Examples of the IT subclass are hammer, nail, saw, screwdriver, wrench, pliers, shovel, rake, lawn mower, lathe, X-ray machine, jack, level, and starting button of a machine. Household appliances are scored in the household class and parts of conveyances are scored in the travel class. (Hall & Van de Castle, 1966, p. 46)

At the other end of the range, words or word groups are coded according to their underlying associative or conceptual communality. For example, "We will *start* the project shortly" and "He was *born* on July second" have a common constituent meaning: beginning" (Laffal, 1968, p. 279). The coder in scoring the units *start* and *born* in the category *begin* (see Table 1) employs the classical coding strategy—the semantic feature *beginning* is marked for words like start and born regardless of the context of their use. Scoring depends only on the capability of the coder to recognize semantically similar words or word groups.

Pragmatic content categories describe some characteristic or condition of the communicator. Thus, instead of locating particular words or word groups with similar meanings the coder "now directly scores resistance, tension, adjustment" (Dittes, 1959, p. 329). For example, Murray's (1956; see Table 1) pragmatic content category, *generalized anxiety*, was defined as follows:

Generalized Anxiety: Included all psychological and somatic expressions of anxiety which are not related to a drive nor related to any specific person or object; general "free floating" anxiety and guilt.

a. "I feel panicky about the thought of death."

b. "I tremble and then it would ease off, then I start again . . . in waves."

c. "I feel tense as if there's some force inside of me trying to get out." (p. 27)

Similarly, White, Fichtenbaum, and Dollard's (1966a, 1966b) pragmatic content category *positive self-evaluation* (see Table 1) would be scored for each sentence judged to reflect, by virtue of its content, a favorable attitude toward the speaker. Thus, "I'm getting a broader idea of myself as an entity and full person" (White et al., 1966a, p. 108) is scored to the category *positive self-evaluation*. This scoring procedure typifies the pragmatic scoring strategy employed by those pragmatic content category systems enumerated in Table 1. As the examples above illustrate, many of the pragmatic content categories of interest to researchers of psychotherapy are used as theoretical constructs by clinicians.

A review of published research suggests that content categories have been most successfully employed to investigate internal psychodynamic processes, motives, drive conditions, characterological traits, and changes in these client characteristics in therapy. The focus on internal psychodynamic processes was made explicit by Murray (1954): "We propose to study the content of verbal behavior in psychotherapy with respect to underlying motives and defenses" (p. 305). Murray (1956) also wrote, "The categories of the content analyses were defined in terms of motivation and conflict, influenced by psychoanalytic and learning theories, and formulated in such a way as to be most relevant to an eventual understanding of the underlying processes of psychotherapy" (p. 23). Auld and Dollard (1966) enumerated the key psychological phenomena that they felt were amenable to investigation with content categories: resistance, transference, unconscious motive, inhibition, dependence, hostility, and interpretation. Dollard and Mowrer (1947), Raimy (1948), Kauffman and Raimy (1949, Murray, Auld, and White (1954), Leary and Gill (1959), Freedman et al. (1951–1952), Lennard and Bernstein (1960, 1969), Auld and White (1959), White et al. (1966a, 1966b), Hall and Van de Castle (1966), and Thibaut and Coules (1952) have carried out investigations that embrace similar assumptions.

As an example of the use of content categories to investigate psychodynamic processes, Murray (1954) found that the frequency of hostility statements in psychotherapy increased as the frequency of defensive statements decreased; he interpreted this as evidence that the client's anxiety had decreased as therapy progressed. Thibaut and Coules (1952), focusing on the relative frequency of hostility statements, reported that residual hostility lessened for subjects who directly communicated their hostile feelings after being provoked. White et al. (1966a) were able to determine that "the therapist focused more than the patient did during the active period of therapy on the areas of sex and evaluation of self" (p. 47) by the high proportion of statements scored to these content categories. By considering sex and evaluation of self as the therapist's target area, the authors were

able to report that the therapist had apparently been successful: The scored content of the patient's talk about these areas increased from the first to last quarter of therapy and was judged to be adaptive. Lennard and Bernstein (1960) reported that in the data collected from four therapists and two patients who interacted in a total of 500 sessions, "the therapists as a group led the patients slightly in the proportion of communication dealing with affect" (p. 85). In comparing the early with the later sessions, they found that "the proportion of both therapist and patient references devoted to feelings increased—almost doubled—for the sample as a whole" (p. 85).

Content categories also reflect personality variables, as well as trends or focal points in the communicated content. Sarason, Ganzer, and Singer (1972) found that high- and low-defensive subjects used different content categories to describe themselves, after having listened to models that differed in self-disclosure style. Kepecs (1977) used content categories to locate the focal conflict of the client. Auld and White (1959) found that experienced therapists are more likely to intervene than are apprentice therapists following a patient's utterance scored as resistance. They also reported that the patient's talk seemed to hang together: Categories on a specific topic were more likely to follow one another than to be followed by a category on a different topic. Although content categories have been used most consistently to investigate internal psychodynamics and characterological traits, one group of researchers (Gottschalk & Frank, 1967; Gottschalk & Gleser, 1969; Gottschalk, Winget, & Gleser, 1969) have concluded that "the relative magnitude of an affect can be validly estimated from the typescript of the speech of an individual using solely content variables and not including paralanguage variables" (Gottschalk & Gleser, 1969, p. 96). A number of studies have been carried out that directly or indirectly address this issue (Cook, 1969; Gottschalk & Frank, 1967; Hart & Brown, 1974; Mahl, 1956, 1959; Markel, Meisels, & Houck, 1964; Markel & Roblin, 1965; Mehrabian & Ferris, 1967). Though the empirical findings are somewhat equivocal, Mahl's (1956) theoretical assessment that "the most valid measures [of transitory states] will be based on the expressive [i.e., extralinguistic] aspects of speech rather than on the manifest content measures" (p. 13) is still most compelling, empirically and theoretically (see Mahl, 1959, for a theoretical discussion of this issue; see Cook, 1969; Markel & Roblin, 1965, for some empirical evidence).

Intersubjective Categories

Intersubjective categories are descriptive of syntactically implied and other relationships between the communicator and recipient. For example, *self-disclosure* implies that the communicator reveals something to the recipient, *question* implies that the communicator seeks information from the recipient, and so forth. (Examples of intersubjective category systems are given in Table 2.) In contrast to content categories, intersubjective categories can typically be defined without reference to the semantic content (or extralinguistic features) of the communication.

Many intersubjective categories can be defined by syntactic features alone (Goodman & Dooley, 1976; Stiles, 1978, in press-b). For example, *question* has a well-attested set of syntactic features associated with it; *self-disclosure* can be defined as a first-person declarative sentence. Since these syntactic features are characteristics of the text, speech units are scored to such categories by means of the classical coding strategy. For example, in Stiles's (1978, in press-a, in press-b) classical intersubjective system (see Table 2), sentences such as "I'd really like to talk about my feelings of being an experimental subject" and "I can't stand needles" would be scored to the category *disclosure form* because of their first-person subjects.

The intersubjective categories used in psychotherapy research are more often descriptive of interpersonal intentions, which may or may not be expressed using the corresponding syntactic form. For example, an utterance identified as a question by syntactic criteria may express the interpersonal intention "asks for information," but it may also express the

interpersonal intention "gives suggestion," as in "Don't you think you should lock the door?"

In Goodman and Dooley's (1976) system, such utterances as "I had the same problem and solved it with . . . " or "Do you think it would work better if you tried . . . " (p. 109) would be scored to the pragmatic content category *advisement* (see Table 2). To judge whether a communicator is giving a suggestion or is asking for information, the coder must infer the communicator's intent and thus employs the pragmatic coding strategy. Pragmatic intersubjective categories are used frequently in interaction research, though they require the coder to make more complicated inferences than their classical counterparts. For instance, *confrontation* is defined by Barnabei, Cormier, and Nye (1974) as "a response indicating some sort of discrepancy in the client's message . . . a 'you said but look' condition" (p. 356). Or, similarly, *direct guidance* (see Table 2) is defined by Strupp and Wallach (1965) as "suggestions for activity either within or outside of the therapeutic framework; giving information, stating an opinion, answering direction questions, speaking as an authority" (p. 118).

In published research, intersubjective categories have been most consistently and successfully employed to measure psychotherapeutic technique, interpersonal roles and relationships in therapy, and therapy process. Intersubjective categories have long been used to describe and teach psychotherapeutic technique. Freud (1912/1958) explicitly argued in favor of using interpretation while he condemned suggestion and disclosure. Rogers (1942, 1951, 1957) advocated reflection (or *restatement* or *clarification*) as a technique. More recently, comprehensive systems of intersubjective categories have been developed to aid in training counselors and therapists (e.g., Goodman & Dooley, 1976; Ivey, 1971). Research evaluating these systems has demonstrated their efficacy in training new professionals (e.g., Moreland, Ivey, & Phillips, 1973).

In view of the different technical prescriptions of the various schools of therapy, it is not surprising that intersubjective categories consistently differentiate the therapeutic interventions made by practitioners of those schools (Auerbach, 1963; Cartwright, 1966; Staples, Sloane, & Whipple, 1976; Strupp, 1955, 1957; Stiles, in press-b). In addition, several studies have used intersubjective categories to describe a single type of psychotherapy (Porter, 1943; Seeman, 1949; Snyder, 1945; Strupp, 1958). In these studies, client-centered, psychoanalytic, existential, gestalt, and behavior therapists have been shown to use characteristic but different profiles of intersubjective categories. Similarly, therapists' style of participation, as well as that of clients, has been related to their choice of intersubjective categories (Rice, 1965, 1973; Rice & Wagstaff, 1967; Segal, 1970). Intersubjective categories have also been employed to characterize the verbal interaction of schizophrenic families. Lennard and Bernstein (1969) found that in a schizophrenic family the child's presentations of self are disconfirmed by the mother and "her presentations (of him) are disconfirmed by him" (p. 125). Bandura, Lipsher, and Miller (1960), Frank and Sweetland (1962), Murray (1956), Rottschafer and Renzaglia (1962), and Winder, Farrukh, Bandura, and Rau (1962) showed that differential utilization of certain intersubjective categories resulted in alteration of the content of the client's speech.

Extralinguistic Categories

Extralinguistic categories are descriptive of vocal noises that do not have the structure of language: modifications such as pitch, resonance, amplitude, and so on of language and other vocal noises, and temporal patternings associated with language behavior. (Examples of extralinguistic category systems are given in Table 3.) Extralinguistic categories are defined without reference to either semantic content or syntactic structure.

Vocal noises that do not have the structure of language have been termed *vocalizations* (Trager, 1958). The following six extralinguistic categories are examples of vocalizations (1–3 from Mahl, 1956; 4–6 from Dibner, 1956, 1958; see Table 3).

1. Whenever the definite *ah* sound occurs, it is scored.

2. An *intruding incoherent sound* is a sound that is absolutely incoherent as a word to the listener.

3. *Stutter.*

4. *Sighing* (or deep breath).

5. *Laughing* includes any kind of laugh or chuckle.

6. *Blocking* occurs when there is groping for the proper expression, indicating unusual hesitation.

Modifications of language and other vocal noises have been termed *qualifiers* (Trager, 1958). Pitch, rhythm, resonance, loudness, intonation, and so on are typical modifications of the noises people emit.

The following three categories (Fairbanks & Pronovost, 1939) illustrate how one specific modification (i.e., pitch) might be delineated for use in analyzing patterns of speech (see Table 3).

1. *Pitch level* is median frequency in cycles per second.

2. *Pitch range* is the highest minus the lowest pitch in cycles per second.

3. *Extent of pitch shifts* is the change in pitch between the last pitch measured in a given phonation and the first pitch measured in the phonation that follows.

Utterance duration, utterance latency, rate of speech production, and so on are observable characteristics of speech production and can be termed *temporal patterning in speech.* An example is Matarazzo, Wiens, Matarazzo, and Saslow's (1968) unit of latency silence (see Table 3), defined as "the duration of time from the moment one person in the dyad terminates an utterance until the second person begins his next comment" (p. 355).

Most extralinguistic categories have used the classical strategy; indeed, investigators who use extralinguistic systems have prided themselves on the objectivity of their systems (e.g., Mahl, 1959; Matarazzo et al., 1968; Phillips et al., 1961; Saslow & Matarazzo, 1959). However, there is no inherent barrier to coding communicator characteristics directly from extralinguistic cues, that is to using the pragmatic strategy for extralinguistic categories. An example of a pragmatic extralinguistic category is Bales's (1950, 1970) category *shows tension* (see Table 2):

> Several varieties of acts are scored in this category, not all of which may seem similar on a superficial level. Laughter, in particular, may seem quite different from signs of anxious emotionality. Signs of anxious emotionality indicate a conflict between acting and withholding action. Minor outbreaks of reactive anxiety may first be mentioned, such as appearing startled, disconcerted, alarmed, dismayed, perturbed, or concerned. Hesitation, speechlessness, flurry, fluster, confusion, trembling, blushing, flushing, stammering, sweating, blocking-up, gulping, swallowing, or wetting the lips persistently may also be included. (Bales, 1970, p. 124)

Judging from published research, extralinguistic categories have been used most successfully to investigate transitory motivational and emotional states. The association of speech disturbances with states of anxiety and tension has been long established by investigators using a variety of speech disturbance categories and indices of anxiety (Dibner, 1956, 1958; Eldred & Price, 1958; Lasswell, 1935; Mahl, 1956, 1959; Panek & Martin, 1959). Dibner (1956), for example, found that situational anxiety produced by the use of ambiguous techniques by the interviewer was associated with speech disturbances in the patient. Kasl and Mahl (1965), Mahl (1959), and Cook (1969) have shown that speech disturbance categories are more sensitive to momentary states of anxiety than to trait anxiety. For example, Cook found that two measures of trait anxiety were not related to the non-ah speech disturbance categories (the measures of anxiety were the Taylor Manifest Anxiety Scale and the McReynolds Assimilation Scale), while transient anxiety was related.

Extralinguistic categories have also been found to be sensitive to other transitory emotional states besides anxiety (Boomer, 1965; Eldred & Price, 1958; Hargreaves, Starkweather, & Blacker, 1965). In fact, the association of affective states with extralinguistic cues is apparently well enough appreciated by most people that the cues can be used to communicate specific emotions. For instance, Fairbanks and Pronovost (1939) have shown that the communication of five

different emotions (i.e., contempt, anger, fear, grief, and indifference) can be reliably distinguished by measures of mean pitch levels, mean pitch ranges, and the mean extent of pitch shifts and that these specific emotions can be reliably identified by listeners. Similarly, Beier and Zautra (1972) have shown that the affective information communicated extralinguistically can even be understood, at least in part, by people of different cultures.

In more recent years, research with extralinguistic categories has identified more subtle—though equally important—transitory states. Boomer and Dittman (1963) suggested that filled pauses might serve as an index of self-monitoring by clients in psychotherapy, and the same function for unfilled pauses has been suggested (Rochester, 1973). Manaugh, Wiens, and Matarazzo (1970) found that subjects instructed to lie to their interviewer showed significant differences in their mean duration of utterance as compared with subjects who were not told to lie. Another group of researchers (Butler, Rice, & Wagstaff, 1962; Duncan, Rice & Butler, 1968; Rice, 1965; Rice & Wagstaff, 1967; Wexler & Butler, 1976) have tentatively identified several constellations of extralinguistic features that have differential associations with therapy outcome, counselor and client participation, and the "good" hour.

Extralinguistic features of speech have also been used to investigate personality traits or characterological makeup. However, reviewers of this field of inquiry have been skeptical of the appropriateness of using extralinguistic categories as a basis for making inferences about personality types. For instance, Starkweather (1961) maintained that "despite the frequent reports of success in identifying personality traits from vocal cues, the numerous [reported] failures . . . leave the writer pessimistic concerning the utility of inferring such traits from non-verbal [but vocal] stimuli" (p. 65). Although recent work in this area has shown some success, parallel studies frequently do not confirm the findings. For example, the personality dimensions of assertiveness/dominance and extraversion/sociability have been identified with high interjudge agreement from extralinguistic features (Markel et al., 1964; Markel & Roblin, 1965; Sherer, 1972). However, in related studies, patterns inconsistent with the above results were obtained (e.g., Hart & Brown, 1974).

Recommendations

Criteria for constructing and evaluating category systems have been proposed by others (Butler et al., 1962; Goodman & Dooley, 1976; Heyns & Zander, 1953; Holsti, 1968, 1969; Lazarsfeld & Barton, 1951; Weick, 1968). These criteria fall into two classes, those that deal with the practicability of systems and those that deal with the internal consistency of categories and category systems. Although our typology can serve as a useful guide in identifying or constructing categories and category systems that conform to the criteria of both classes, we limit our discussion to (a) suggesting means by which a researcher can attain internally consistent categories and category systems and (b) pointing out some common methodological problems. (See Goodman & Dooley, 1976, for a recent set of practicability criteria.)

System Consistency Criteria

1. The categories within the system should be mutually exclusive, that is, "there should be one and only one place to put an item within a given classification system" (Lazarsfeld & Barton, 1951, p. 151; see also Butler et al., 1962; Holsti, 1969).

2. The categories within the system should be exhaustive: All relevant items in the sample must be capable of being placed into a category (Holsti, 1969).

3. The categories within the system should be derivable from a single classification principle; that is, "conceptually different levels of analysis must be kept separate" (Holsti, 1969, p. 100). "When an object is classified at the same time from more than one aspect, each aspect must have its own separate set of categories" (Lazarsfeld & Barton, 1951, p. 157); that is, if one is interested in classifying items in terms of a number of different aspects simultaneously, a fully multidimensional classification must be set up.

Our typology suggests two recommendations for meeting the system consistency criteria: (a) Categories should be pure types, that is, content, intersubjective, or extralinguistic combined with either the classical or the pragmatic strategy, rather than conjunctive or disjunctive mixtures, and (b) categories within a system (or subsystem) should be of the same type.

Thus, we identify two types of problematic coding schemes: categories that are conjunctive or disjunctive mixtures and systems that mix together category types. For example, the Strupp (1957) category *structuring* (see Table 2) is scored when the therapist is judged to be structuring the therapeutic situation (i.e., an intersubjective category type) or when the therapist is discussing theory (i.e., a content category type). A mixed system is one in which the categories are of more than one type. For example, Heyns and Zander (1953, p. 391) pointed out that Bales's (1950) interaction process analysis system (see Table 2) contains more than one type of category: "Category 3, Shows Solidarity and Category 2, Shows Tension Release seem to be descriptions of interaction along affective dimensions" (and are defined with reference to extralinguistic features); "Category 5, Gives Opinion, Category 6, Gives Orientation and Category 4, Gives Suggestion refer to intellectual problem-solving activity of the group" (and are defined with reference to intersubjective features).

Following the recommendations suggested by the typology does not constrain the researcher from employing a variety of category and category system construction strategies, but does help facilitate the systematic organization of such strategies. For instance, the sentence "Did m-mother really leave?" can be coded to any of the three category types, using either the classical or the pragmatic strategy. It might be scored to the classical content category *mother*, the pragmatic content category *separation anxiety*, the classical intersubjective category *question*, the pragmatic intersubjective category *seeking reassurance*, the classical extralinguistic category *stutter*, and the pragmatic extralinguistic category *nervousness*. Insisting that a system contain only one type of category (i.e., recommendation b) and that categories should be pure types (i.e., recommendation a) does not guarantee that categories within a system will be mutually exclusive (i.e., system consistency criterion 1) or that each of the categories will be derivable from a single classification principle (i.e., system consistency criterion 3), but permitting a system to contain more than one category type or mixed categories virtully guarantees that the system will not be mutually exclusive or derivable from a single principle.

Application of the typology makes category identification explicit and reasonably simple, permitting researchers to separate conceptually and empirically different levels of analysis that heretofore have too often been lumped together. Thus, logically sound category systems, consisting of pure category types, can be constructed with reference to the three sets of features, content, intersubjective, and extralinguistic. By keeping these dimensions separate, researchers interested in more than one set of features can build truly multidimensional systems.

Our view suggests that the content, intersubjective, and extralinguistic categories correspond to distinct channels of verbal communication, which convey different types of psychological information. Information concerning the speaker's personality structure and dynamics is carried primarily in the content channel; information concerning the nature of the speaker's current relationship to the other person is carried primarily in the intersubjective channel; and information concerning the speaker's transitory emotional state is carried primarily in the extralinguistic channel. Although these associations are not exclusive, they are strong enough to recommend that investigators interested in one type of information would be wise to select a system that codes the corresponding channel.

The division of the study of language behavior in psychotherapy into three areas reflects not only trends in the empirical studies reviewed and theoretical and methodological considerations but also historical and philosophical views at large. For example, Johnson (1976), influenced by Kuhn's (1970) analysis of the structure of scientific enterprises, argued "that two of the major under-

lying paradigms in present day psychology are the behavioristic paradigm and the Freudian paradigm" (p. 4). He pointed out that those psychotherapy process researchers "who have been influenced by Freud have stressed the importance of the content of the interview" (p. 4), while those who have been heavily influenced by behaviorism (or more specifically, positivism) have "focused on clearly denotable subject behaviors" (p. 5), as are measured by the Interaction Chronograph (Matarazzo et al., 1968). A third paradigm in psychotherapy process research has grown out of the search for variables with "systematic interpersonal reference" (Freedman et al., 1951–1952, p. 143). Researchers influenced by this interpersonal orientation describe relationships in terms of the sorts of information conveyed through intersubjective categories.

If from one perspective process studies of psychotherapy have appeared chaotic, repetitive, and so forth, we have found that the typology provides a descriptive framework within which many of the unresolved and seemingly unrelated problems tend to cluster and come into clearer focus. We hope that this framework will provide some of the impetus needed to begin to move psychotherapy process research out of its first stage.

References

Auerbach, A. H. An application of Strupp's method of content analysis to psychotherapy. *Psychiatry,* 1963, *26,* 137–148.

Auld, F., Jr., & Dollard, J. Measurement of motivational variables in psychotherapy. In L. A. Gottschalk & A. H. Auerbach (Eds.), *Methods of research in psychotherapy.* New York: Appleton-Century-Crofts, 1966.

Auld, F., Jr., & Murray, E. J. Content-analysis studies of psychotherapy. *Psychological Bulletin,* 1955, *52,* 377–395.

Auld, F., Jr., & White, A. M. Sequential dependencies in psychotherapy. *Journal of Abnormal and Social Psychology,* 1959, *58,* 100–104.

Bales, R. F. *Interaction process analysis.* Reading, Mass.: Addison-Wesley, 1950.

Bales, R. F. *Personality and interpersonal behavior.* New York: Holt, Rinehart & Winston, 1970.

Bandura, A., Lipsher, D. H., & Miller, P. E. Psychotherapists' approach–avoidance reactions to patients' expressions of hostility. *Journal of Consulting Psychology,* 1960, *24,* 1–8.

Barnabei, F., Cormier, W. H., & Nye, L. S. Determining the effects of three counselor verbal responses on client verbal behavior. *Journal of Counseling Psychology,* 1974, *21,* 355–359.

Beier, E. G., & Zautra, A. J. Identification of vocal communication of emotions across cultures. *Journal of Consulting and Clinical Psychology,* 1972, *39,* 166.

Berelson, B. *Content analysis in communication research.* Glencoe, Ill.: Free Press, 1952.

Bieber, M. R., Patton, M. J., & Fuhriman, A. A metalanguage analysis of counselor and client verb usage in counseling. *Journal of Counseling Psychology,* 1977, *24,* 264–271.

Boomer, D. S. Hesitation and grammatical encoding. *Language and Speech,* 1965, *8,* 148–158.

Boomer, D. S., & Dittman, A. T. Speech rate, filled pause, and body movement in interviews. *Journal of Nervous and Mental Disease,* 1963, 7, 324–327.

Butler, J. M., Rice, L. N., & Wagstaff, A. K. On the naturalistic definition of variables: An analogue of clinical analysis. In H. H. Strupp & L. Luborsky (Eds.), *Research in psychotherapy* (Vol. 2). Washington, D.C.: American Psychological Association, 1962.

Cartwright, R. D. A comparison of the response to psychoanalytic and client-centered psychotherapy. In L. A. Gottschalk & A. H. Auerbach (Eds.), *Methods of research in psychotherapy.* New York: Appleton-Century-Crofts, 1966.

Cook, M. Anxiety, speech disturbance and speech rate. *British Journal of Clinical Psychology,* 1969, *8,* 13–21.

Dibner, A. S. Cue counting: A measure of anxiety in interviews. *Journal of Consulting Psychology,* 1956, *20,* 475–478.

Dibner, A. S. Ambiguity and anxiety. *Journal of Abnormal and Social Psychology,* 1958, *56,* 165–174.

Dittes, J. E. Previous studies bearing on content analysis of psychotherapy. In J. Dollard & F. Auld, Jr. (Eds.), *Scoring human motives: A manual.* New Haven, Conn.: Yale University Press, 1959.

Dollard, J., & Auld, F., Jr. *Scoring human motives: A manual.* New Haven, Conn.: Yale University Press, 1959.

Dollard, J., & Mowrer, D. H. A method of measuring tension in written documents. *Journal of Abnormal and Social Psychology,* 1947, *42,* 3–32.

Duncan, S., Jr., Rice, L. N., & Butler, J. M. Therapists' paralanguage in peak and poor psychotherapy hours. *Journal of Abnormal Psychology,* 1968, *73,* 566–570.

Eldred, S. H., & Price, D. B. The linguistic evaluation of feeling states in psychotherapy. *Psychiatry,* 1958, *21,* 115–121.

Fairbanks, G., & Pronovost, W. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs,* 1939, *6,* 87–104.

Frank, G. H., & Sweetland, A. A study of the process of psychotherapy: The verbal interaction.

Journal of Consulting Psychology, 1962, *26,* 135–138.

Freedman, M. B., Leary, T. F., Ossorio, A. G., & Coffey, H. S. The interpersonal dimension of personality. *Journal of Personality,* 1951–1952, *20,* 143–161.

Freud, S. [Recommendations to physicians practicing psychoanalysis.] In J. Strachey (Ed. and trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 12). London: Hogarth Press, 1958. (Originally published, 1912.)

Gerbner, G., Holsti, O. R., Krippendorf, K., Paisley, W. J., & Stone, P. J. (Eds.). *The analysis of communication content: Developments in scientific theories and computer techniques.* New York: Wiley, 1969.

Goodman, G., & Dooley, D. A framework for help-intended interpersonal communication. *Psychotherapy: Theory, Research, and Practice,* 1976, *13,* 106–117.

Gottschalk, L. A., & Frank, E. C. Estimating the magnitude of anxiety from speech. *Behavioral Science,* 1967, *12,* 289–295.

Gottschalk, L. A., & Gleser, G. D. *The measurement of psychological states through the content analysis of verbal behavior.* Berkeley: University of California Press, 1969.

Gottschalk, L. A., Winget, C. N., & Gleser, G. D. *Manual of instructions for using the Gottschalk-Gleser content analysis scales: Anxiety, hostility, and social aberration-personal disorganization.* Berkeley: University of California Press, 1969.

Hall, C. S., & Van de Castle, R. L. *The content analysis of dreams.* New York: Appleton-Century-Crofts, 1966.

Hargreaves, W. A., Starkweather, J. A., & Blacker, K. H. Voice quality in depression. *Journal of Abnormal Psychology,* 1965, *70,* 218–220.

Hart, R. J., & Brown, B. L. Interpersonal information: Information conveyed by the content and vocal aspects of speech. *Speech Monographs,* 1974, *41,* 371–380.

Heyns, R. W., & Zander, A. F. Observation of group behavior. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences.* New York: Dryden Press, 1953.

Holsti, O. R. Content analysis. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed.). Reading, Mass.: Addison-Wesley, 1968.

Holsti, O. R. *Content analysis for the social sciences and humanities.* Reading, Mass.: Addison-Wesley, 1969.

Ivey, A. E. *Microcounseling: Innovations in interviewing training.* Springfield, Ill.: Charles C Thomas, 1971.

Johnson, R. F. Q. Pitfalls in research: The interview as an illustrative model. *Psychological Reports,* 1976, *38,* 3–17.

Kaplan, A. Content analysis and the theory of signs. *Philosophy of Science,* 1943, *10,* 230–247.

Kasl, S. V., & Mahl, G. F. The relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology,* 1965, *1,* 425–433.

Kauffman, P. E., & Raimy, V. C. Two methods of assessing therapeutic progress. *Journal of Abnormal and Social Psychology,* 1949, *44,* 379–385.

Kepecs, J. G. Teaching psychotherapy by use of brief typescripts. *American Journal of Psychotherapy,* 1977, *31,* 383–393.

Kiesler, D. J. *The process of psychotherapy.* Chicago: Aldine, 1973.

Klein, G. H., Mathieu, P. L., Gendlin, E. T., & Kiesler, D. J. *The experiencing scale: A research and training manual* (2 vols.). Madison: Wisconsin Psychiatric Institute, Bureau of Audio Visual Instruction, 1970.

Kuhn, T. S. *The structure of scientific revolutions* (2nd ed.) Chicago: University of Chicago Press, 1970.

Labov, W., & Fanshel, D. *Therapeutic discourse: Psychotherapy as conversation.* New York: Academic Press, 1977.

Laffal, J. An approach to the total content analysis of speech in psychotherapy. In J. M. Shlien (Ed.), *Research in psychotherapy* (Vol. 3). Washington, D.C.: American Psychological Association, 1968.

Lasswell, H. D. Verbal references and physiological changes during the psychoanalytic interview: A preliminary communication. *Psychoanalytic Review,* 1935, *22,* 10–24.

Lazarsfeld, P. F., & Barton, A. H. Qualitative measurement in the social sciences: Classification, typologies, and indices. In D. Lerner & H. D. Lasswell (Eds.), *The policy sciences: Recent developments in scope and method.* Stanford, Calif.: Stanford University Press, 1951.

Leary, T., & Gill, M. The dimensions and a measure of the process of psychotherapy: A system for the analysis of the content of clinical evaluations and patient–therapist verbalizations. In E. A. Rubinstein & M. B. Parloff (Eds.), *Research in psychotherapy* (Vol. 1). Washington, D.C.: American Psychological Association, 1959.

Lennard, H. L., & Bernstein, A. *The anatomy of psychotherapy: Systems of communication and expectation.* New York: Columbia University Press, 1960.

Lennard, H. L., & Bernstein, A. *Patterns in human interaction.* San Francisco: Jossey-Bass, 1969.

Mahl, G. F. Disturbances and silences in the patient's speech in psychotherapy. *Journal of Abnormal and Social Psychology,* 1956, *53,* 1–15.

Mahl, G. F. Exploring emotional states by content analysis. In I. Pool (Ed.), *Trends in content analysis.* Urbana: University of Illinois Press, 1959.

Mahl, G. F., & Schulze, G. Psychological research in the extralinguistic area. In T. A. Sebeok, A. S. Hayes, & M. C. Bateson (Eds.), *Approaches to semiotics.* The Hague, Netherlands: Mouton, 1964.

Manaugh, T. S., Wiens, A. N., & Matarazzo, J. D. Content saliency and interview speech behavior. *Journal of Clinical Psychology,* 1970, *26,* 17–24.

Markel, N. N., Meisels, M., & Houck, J. E. Judging personality from voice quality. *Journal of Abnormal and Social Psychology,* 1964, *69,* 458–463.

Markel, N. N., & Roblin. The effect of content and sex-of-judge on judgments of personality from voice. *International Journal of Social Psychiatry,* 1965, *11,* 295–300.

Marsden, G. Content-analysis studies of therapeutic interviews: 1954 to 1964. *Psychological Bulletin,* 1965, *63,* 298–321.

Marsden, G. Content analysis studies of psychotherapy: 1954 to 1968. In A. E. Gergins & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change.* New York: Wiley, 1971.

Matarazzo, J. D., & Wiens, A. N. Speech behavior as an objective correlate of empathy and outcome in interview and psychotherapy research. *Behavior Modification,* 1977, *1,* 453–480.

Matarazzo, J. D., Wiens, A. N., Matarazzo, R. G., & Saslow, G. Speech and silence behavior in clinical psychotherapy and its laboratory correlates. In J. M. Schlien, H. F. Hunt, J. D. Matarazzo, & C. Savage (Eds.), *Research in psychotherapy* (Vol. 3). Washington, D.C.: American Psychological Association, 1968.

Mehrabian, A. *Nonverbal communication.* Chicago: Aldine-Atherton, 1972.

Mehrabian, A., & Ferris, S. R. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology,* 1967, *31,* 248–252.

Meltzoff, J., & Kornreich, M. *Research in psychotherapy.* New York: Atherton Press, 1970.

Moreland, J. R., Ivey, A. E., & Phillips, J. An evaluation of microcounseling as an interviewer training tool. *Journal of Consulting and Clinical Psychology,* 1973, *41,* 294–300.

Murray, E. J. A case study in a behavioral analysis of psychotherapy. *Journal of Abnormal and Social Psychology,* 1954, *49,* 305–310.

Murray, E. J. A content-analysis method for studying psychotherapy. *Psychological Monographs,* 1956, *70*(13, Whole No. 420).

Murray, E. J., Auld, R., Jr., & White, A. M. A psychotherapy case showing progress but no decrease in the discomfort-relief quotient. *Journal of Consulting Psychology,* 1954, *18,* 349–353.

Panek, D. M., & Martin, B. The relationship between GSR and speech disturbance in psychotherapy. *Journal of Abnormal and Social Psychology,* 1959, *58,* 402–405.

Patton, M. J., Fuhriman, A., & Bieber, M. R. A model and a metalanguage for research on psychological counseling. *Journal of Counseling Psychology,* 1977, *24,* 25–34.

Phillips, J. S., Matarazzo, R. G., Matarazzo, J. R., Saslow, G., & Kanfer, F. H. Relationships between descriptive and interaction behavior in interviews. *Journal of Consulting Psychology,* 1961, *25,* 260–266

Pool, I. (Ed.). *Trends in content analysis.* Urbana: University of Illinois Press, 1959.

Porter, E. H., Jr. The development and evaluation of a measure of counseling interview procedures. *Educational and Psychological Measurement,* 1943, *3,* 105–126.

Raimy, V. C. Self reference in counseling interviews. *Journal of Consulting Psychology,* 1948, *12,* 153–163.

Reusch, J., & Bateson, G. Structure and process in social relations. *Psychiatry,* 1949, *12,* 105–124.

Rice, L. N. Therapists' style of participation and case outcome. *Journal of Consulting Psychology,* 1965, *29,* 155–160.

Rice, L. N. Client behavior as a function of therapist style and client resources. *Journal of Counseling Psychology,* 1973, *20,* 306–331.

Rice, L. N., & Wagstaff, A. K. Client voice quality and expressive style as indexes of productive psychotherapy. *Journal of Consulting Psychology,* 1967, *31,* 557–563.

Rochester, S. R. The significance of pauses in spontaneous speech. *Journal of Psycholinguistics Research,* 1973, *2,* 51–81.

Rogers, C. R. *Counseling and psychotherapy.* Boston: Houghton Mifflin, 1942.

Rogers, C. R. *Client-centered therapy.* Boston: Houghton Mifflin, 1951.

Rogers, C. R. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology,* 1957, *21,* 95–103.

Rogers, C. R. A process conception of psychotherapy. *American Psychologist,* 1958, *13,* 142–149.

Rogers, C. R. A tentative scale for the measurement of process in psychotherapy. In E. A. Rubinstein & M. B. Parloff (Eds.), *Research in psychotherapy* (Vol. 1). Washington, D.C.: American Psychological Association, 1959.

Rottschafer, R. H., & Renzaglia, G. A. The relationship of dependent-like verbal behaviors to counselor style and induced set. *Journal of Consulting Psychology,* 1962, *26,* 172–177.

Russell, R. L. Speech acts, conversational sequencing, and rules: A symposium review of Labov and Fanshel's *Therapeutic discourse. Contemporary Sociology,* in press.

Sarason, I. G., Ganzer, V. J., & Singer, M. Effects of modeled self-disclosure on the verbal behavior of persons differing in defensiveness. *Journal of Consulting and Clinical Psychology,* 1972, *39,* 483–490.

Saslow, G., & Matarazzo, J. D. A technique for studying changes in interview behavior. In E. A. Rubinstein & M. B. Parloff (Eds.), *Research in psychotherapy* (Vol. 1). Washington, D.C.: American Psychological Association, 1959.

Seeman, J. A study of the process of nondirective therapy. *Journal of Consulting Psychology,* 1949, *13,* 157–168.

Segal, B. A-B distinction and therapeutic interaction. *Journal of Consulting and Clinical Psychology,* 1970, *34,* 442–446.

Sherer, K. R. Judging personality from voice: A cross-cultural approach to an old issue in interpersonal perception. *Journal of Personality,* 1972, *40,* 191–210.

Snyder, W. U. An investigation of the nature of nondirective psychotherapy. *Journal of General Psychology,* 1945, *33,* 193–223.
Snyder, W. U. *Dependency in psychotherapy: A casebook.* New York: Macmillan, 1963.
Staples, F. R., Sloane, R. B., & Whipple, K. Process and outcome in psychotherapy and behavior therapy. *Journal of Consulting and Clinical Psychology,* 1976, *44,* 340–350.
Starkweather, J. A. Vocal communication of personality and human feelings. *Journal of Communication,* 1961, *11,* 63–72.
Stiles, W. B. Verbal response modes and dimensions of interpersonal roles: A method of discourse analysis. *Journal of Personality and Social Psychology,* 1978, *36,* 693–703.
Stiles, W. B. *Manual for a taxonomy of verbal response modes.* Chapel Hill: University of North Carolina, Institute for Research in Social Science, in press. (a)
Stiles, W. B. Verbal response modes and psychotherapeutic technique. *Psychiatry,* in press. (b)
Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. *The general inquirer: A computer approach to content analysis.* Cambridge, Mass.: M.I.T. Press, 1966.
Strupp, H. H. An objective comparison of Rogerian and psychoanalytic techniques. *Journal of Consulting Psychology,* 1955, *19,* 1–7.
Strupp, H. H. A multidimensional system for analyzing psychotherapeutic techniques. *Psychiatry,* 1957, *20,* 293–312.
Strupp, H. H. The performance of psychoanalytic and client-centered therapists in an initial interview. *Journal of Consulting Psychology,* 1958, *22,* 265–274.
Strupp, H. H., & Wallach, M. S. A further study of psychiatrists' responses in quasitherapy situations. *Behavioral Science,* 1965, *10,* 113–134.
Thibaut, J. W., & Coules, J. The role of communication in the reduction of interpersonal hostility. *Journal of Abnormal and Social Psychology,* 1952, *47,* 770–777.
Trager, G. L. Paralanguage: A first approximation. *Studies in Linguistics,* 1958, *13,* 1–12.
Weick, K. E. Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (2nd ed.). Reading, Mass.: Addison-Wesley, 1968.
Wexler, D. A., & Butler, J. M. Therapist modification of client expressiveness in client-centered therapy. *Journal of Counseling Psychology,* 1976, *44,* 261–265.
White, A. M., Fichtenbaum, L., & Dollard, J. A. A content measure of changes attributable to psychotherapy. *American Journal of Orthopsychiatry,* 1966, *36,* 41–49. (a)
White, A. M., Fichtenbaum, L., & Dollard, J. A. Measuring change: A verbal and non-verbal content analysis method. *Psychotherapy,* 1966, *3,* 107–113. (b)
Winder, C. L., Farrukh, A. Z., Bandura, A., & Rau, L. C. Dependency of patients, psychotherapists' responses, and aspects of psychotherapy. *Journal of Consulting Psychology,* 1962, *26,* 129–134.

Received January 3, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 2, 420–428

Intraclass Correlations: Uses in Assessing Rater Reliability

Patrick E. Shrout and Joseph L. Fleiss
Division of Biostatistics
Columbia University, School of Public Health

Reliability coefficients often take the form of intraclass correlation coefficients. In this article, guidelines are given for choosing among six different forms of the intraclass correlation for reliability studies in which n targets are rated by k judges. Relevant to the choice of the coefficient are the appropriate statistical model for the reliability study and the applications to be made of the reliability results. Confidence intervals for each of the forms are reviewed.

Most measurements in the behavioral sciences involve measurement error, but judgments made by humans are especially plagued by this problem. Since measurement error can seriously affect statistical analysis and interpretation, it is important to assess the amount of such error by calculating a reliability index. Many of the reliability indices available can be viewed as versions of the intraclass correlation, typically a ratio of the variance of interest over the sum of the variance of interest plus error (Bartko, 1966; Ebel, 1951; Haggard, 1958).

There are numerous versions of the intraclass correlation coefficient (*ICC*) that can give quite different results when applied to the same data. Unfortunately, many researchers are not aware of the differences between the forms, and those who are often fail to report which form they used. Each form is appropriate for specific situations defined by the experimental design and the conceptual intent of the study. Unfortunately, most textbooks (e.g., Hayes, 1973; Snedecor & Cochran, 1967; Winer, 1971) describe only one or two forms of the several possible. Making the plight of the researchers worse, some of the older references (e.g., Haggard, 1958) contain mistakes that have been corrected in a variety of forums (Bartko, 1966; Feldt, 1965).

In this article, we attempt to give a set of guidelines for researchers who have use for intraclass correlations. Six forms of the *ICC* are discussed here. We discuss these forms in the context of a reliability study of the ratings of several judges. This context is a special case of the one-facet generalizability study (G study) discussed by Cronbach, Gleser, Nanda, and Rajaratnam (1972). The results we present are applicable to other one-facet studies, but we find the case of judges most compelling.

The guidelines for choosing the appropriate form of the *ICC* call for three decisions: (a) Is a one-way or two-way analysis of variance (ANOVA) appropriate for the analysis of the reliability study? (b) Are differences between the judges' mean ratings relevant to the reliability of interest? (c) Is the unit of analysis an individual rating or the mean of several ratings? The first and second decisions pertain to the appropriate statistical model for the reliability study, and the second and the third to the potential use of its results.

Models for Reliability Studies

In a typical interrater reliability study, each of a random sample of n targets is rated independently by k judges. Three different

This work was supported in part by Grant 1 R01 MH 28655-01A1 PCR from the National Institute of Mental Health.

Requests for reprints should be sent to Patrick E. Shrout, Division of Biostatistics, Columbia University, School of Public Health, 600 West 168th Street, New York, New York 10032.

Table 1
Analysis of Variance and Mean Square Expectations for One- and Two-Way Random Effects and Two-Way Mixed Model Designs

Source of variation	df	MS	*EMS*: One-way random effects for Case 1	Two-way random effects for Case 2	Two-way mixed model for Case 3[a]
Between targets	$n - 1$	BMS	$k\sigma_T^2 + \sigma_W^2$	$k\sigma_T^2 + \sigma_I^2 + \sigma_E^2$	$k\sigma_T^2 + \sigma_E^2$
Within target	$n(k - 1)$	WMS	σ_W^2	$\sigma_J^2 + \sigma_I^2 + \sigma_E^2$	$\theta_J^2 + f\sigma_I^2 + \sigma_E^2$
Between judges	$(k - 1)$	JMS	—	$n\sigma_J^2 + \sigma_I^2 + \sigma_E^2$	$n\theta_J^2 + f\sigma_I^2 + \sigma_E^2$
Residual	$(n - 1)(k - 1)$	EMS	—	$\sigma_I^2 + \sigma_E^2$	$f\sigma_I^2 + \sigma_E^2$

[a] $f = k/(k - 1)$ for the last three entries in this column.

cases of this kind of study can be defined:

1. Each target is rated by a different set of k judges, randomly selected from a larger population of judges.

2. A random sample of k judges is selected from a larger population, and each judge rates each target, that is, each judge rates n targets altogether.

3. Each target is rated by each of the same k judges, who are the only judges of interest.

Each kind of study requires a separately specified mathematical model to describe its results. The models each specify the decomposition of a rating made by the ith judge on the jth target in terms of various effects. Among the possible effects are those for the ith judge, for the jth target, for the interaction between judge and target, for the constant level of ratings, and for a random error component. Depending on the way the study is designed, different ones of these effects are estimable, different assumptions must be made about the estimable effects, and the structure of the corresponding ANOVA will be different. The various models that result from the above cases correspond to the standard ANOVA models, as discussed in a text such as Hayes (1973). We review these models briefly below.

Under Case 1, the effects due to judges, to the interaction between judge and target, and to random error are not separable. Let x_{ij} denote the ith rating $(i = 1, \ldots, k)$ on the jth target $(j = 1, \ldots, n)$. For Case 1, we assume the following linear model for x_{ij}:

$$x_{ij} = \mu + b_j + w_{ij}. \quad (1)$$

In this equation, the component μ is the overall population mean of the ratings; b_j is the difference from μ of the jth target's so-called true score (i.e., the mean across many repeated ratings on the jth target); and w_{ij} is a residual component equal to the sum of the inseparable effects of the judge, the Judge × Target interaction, and the error term. The component b_j is assumed to vary normally with a mean of zero and a variance of σ_T^2 and to be independent of all other components in the model. It is also assumed that the w_{ij} terms are distributed independently and normally with a mean of zero and a variance of σ_W^2. The expected mean squares in the ANOVA table appropriate to this kind of study (technically a one-way random effects layout) appear under Case 1 in Table 1.

The models for Case 2 and Case 3 differ from the model for Case 1 in that the components of w_{ij} are further specified. Since the same k judges rate all n targets, the component representing the ith judge's effect may be estimated. The equation

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij} \quad (2)$$

is appropriate for both Case 2 and Case 3. In Equation 2, the terms x_{ij}, μ, and b_j are defined as in Equation 1; a_i is the difference from μ of the mean of the ith judge's ratings; $(ab)_{ij}$ is the degree to which the ith judge departs from his or her usual rating tendencies when confronted by the jth target; and e_{ij} is the random error in the ith judge's scoring of the jth target. In both Cases 2 and 3 the target component b_j is assumed to vary normally with a mean of zero and variance σ_T^2 (as in

Case 1), and the error terms e_{ij} are assumed to be independently and normally distributed with a mean of zero and variance σ_e^2.

Case 2 differs from Case 3, however, with regard to the assumptions made concerning a_i and $(ab)_{ij}$ in Equation 2. Under Case 2, a_i is a random variable that is assumed to be normally distributed with a mean of zero and variance σ_J^2; under Case 3, it is a fixed effect subject to the constraint $\Sigma a_i = 0$. The parameter corresponding to σ_J^2 is $\theta_J^2 = \Sigma a_i^2/(k-1)$.

In the absence of repeated ratings by each judge on each target, the components $(ab)_{ij}$ and e_{ij} cannot be estimated separately. Nevertheless, they must be kept separate in Equation 2 because the properties of the interaction are different in the two cases being considered. Under Case 2, all the components $(ab)_{ij}$, where $i = 1, \ldots, k; j = 1, \ldots, n$, can be assumed to be mutually independent with a mean of zero and variance σ_I^2. Under Case 3, however, independence can only be assumed for interaction components that involve different targets. For the same target, say the jth, the components are assumed to satisfy the constraint

$$\sum_{i=1}^{k} (ab)_{ij} = 0.$$

A consequence of this constraint is that any two interaction components for the same target, say $(ab)_{ij}$ and $(ab)_{i'j}$, are *negatively* correlated (see, e.g., Scheffé, 1959, section 8.1). The reason is that because of the above constraint,

$$0 = \operatorname{var}\left[\sum_{i=1}^{k} (ab)_{ij}\right] = k \operatorname{var}\left[(ab)_{ij}\right] + k(k-1)\operatorname{cov}\left[(ab)_{ij}, (ab)_{i'j}\right] = k\sigma_I^2 + k(k-1)c,$$

say, where c is the common covariance between interaction effects on the same target. Thus

$$c = \frac{-\sigma_I^2}{k-1}. \qquad (3)$$

The expected mean squares in the ANOVA for Case 2 (technically a two-way random effects layout) and Case 3 (technically a two-way mixed effects layout) are shown in the final two columns of Table 1. The differences are that the component of variance due to the interaction (σ_I^2) contributes additively to each expectation under Case 2, whereas under Case 3, it does not contribute to the expected mean square between targets, and it contributes additively to the other expectations after multiplication by the factor $f = k/(k-1)$.

In the remainder of this article, various intraclass correlation coefficients are defined and estimated. A rigorous definition is adopted for the *ICC*, namely, that the *ICC* is the correlation between one measurement (either a single rating or a mean of several ratings) on a target and another measurement obtained on that target. The *ICC* is thus a bona fide correlation coefficient that, as is shown below, is often but not necessarily identical to the component of variance due to targets divided by the sum of it and other variance components. In fact, under Case 3, it is possible for the population value of the *ICC* to be negative (a phenomenon pointed out some years ago by Sitgreaves [1960]).

Decision 1: A One- or Two-Way Analysis of Variance

In selecting the appropriate form of the *ICC*, the first step is the specification of the appropriate statistical model for the reliability study (or G study). Whether one analyzes the data using a one-way or a two-way ANOVA depends on whether the study is designed according to Case 1, as described earlier, or according to Case 2 or 3. Under Case 1, the one-way ANOVA yields a between-targets mean square (*BMS*) and a within-target mean square (*WMS*).

From the expectations of the mean squares shown for Case 1 in Table 1, one can see that *WMS* is as unbiased estimate of σ_W^2; in addition, it is possible to get an unbiased estimate of the target variance σ_T^2 by subtracting *WMS* from *BMS* and dividing the difference by the number of judges per target. Since the w_{ij} terms in the model for Case 1 (see Equation 1) are assumed to be independent, one can see that σ_T^2 is equal to the covariance between two ratings on a target. Using this information, one can write a formula to estimate ρ, the population value of the *ICC* for Case 1. Because the covariance of the ratings is a variance term, the index

in this case takes the form of a variance ratio:

$$\rho = \sigma_T^2/(\sigma_T^2 + \sigma_W^2).$$

The estimate, then, takes the form

$$ICC(1, 1) = \frac{BMS - WMS}{BMS + (k-1)WMS},$$

where k is the number of judges rating each target. It should be borne in mind that while $ICC(1, 1)$ is a consistent estimate of ρ, it is biased (cf. Olkin & Pratt, 1958).

If the reliability study has the design of Case 2 or 3, a Target $\times$ Judges two-way ANOVA is the appropriate mode of analysis. This analysis partitions the within-target sum of squares into a between-judges sum of squares and a residual sum of squares. The corresponding mean squares in Table 1 are denoted JMS and EMS.

It is crucial to note that the expectation of BMS under Cases 2 and 3 is different from that under Case 1, even though the computation of this term is the same. Because the effect of judges is the same for all targets under Cases 2 and 3, interjudge variability does not affect the expectation of BMS. An important practical implication is that for a given population of targets, the observed value of BMS in a Case 1 design tends to be larger than that in a Case 2 or Case 3 design.

There are important differences between the models for Case 2 and Case 3. Consider Case 2 first. From Table 1 one can see that an estimate of the target variance σ_T^2 can be obtained by subtracting EMS from BMS and dividing the difference by k. Under the assumptions of Case 2 that judges are randomly sampled, the covariance between two ratings on a target is again σ_T^2, and the expression for

Table 2
Four Ratings on Six Targets

	Judge			
Target	1	2	3	4
1	9	2	5	8
2	6	1	3	2
3	8	4	6	8
4	7	1	2	6
5	10	5	6	9
6	6	2	4	7

Table 3
Analysis of Variance for Ratings

Source of variance	*df*	*MS*
Between targets	5	11.24
Within target	18	6.26
Between judges	3	32.49
Residual	15	1.02

the parameter ρ is again a variance ratio:

$$\rho = \sigma_T^2/(\sigma_T^2 + \sigma_J^2 + \sigma_I^2 + \sigma_E^2).$$

It is estimated by

$$ICC(2, 1) = \frac{BMS - EMS}{BMS + (k-1)EMS + k(JMS - EMS)/n},$$

where n is the number of targets. To our knowledge, Rajaratnam (1960) and Bartko (1966) were the first to give this form. Like $ICC(1, 1)$, $ICC(2, 1)$ is a biased but consistent estimator of ρ.

As we have discussed, the statistical model for Case 3 differs from Case 2 because of the assumption that judges are fixed. As the reader can verify from Table 1, one implication of this is that no unbiased estimator of σ_T^2 is available when $\sigma_I^2 > 0$. On the other hand, under Case 3, σ_T^2 is no longer equal to the covariance between ratings on a target, because of the correlated interaction terms in Equation 2. Because the interaction terms on the same target are correlated, as shown in Equation 3, the actual covariance is equal to $\sigma_T^2 - \sigma_I^2/(k-1)$. Another implication of the Case 3 assumption is that the total variance is equal to $\sigma_T^2 + \sigma_I^2 + \sigma_E^2$, and thus the correlation is

$$\rho = \frac{\sigma_T^2 - \sigma_I^2/(k-1)}{\sigma_T^2 + \sigma_I^2 + \sigma_E^2}.$$

This is estimated consistently but with bias by

$$ICC(3, 1) = \frac{BMS - EMS}{BMS + (k-1)EMS}.$$

As is discussed in the next section, the interpretation of $ICC(3, 1)$ is quite different from that of $ICC(2, 1)$.

It is not likely that $ICC(2, 1)$ or $ICC(3, 1)$ will ever be erroneously used in a Case 1 study, since the appropriate mean squares would not be available. The misuse of $ICC(1, 1)$ on data

Table 4
Correlation Estimates From Six Intraclass Correlation Forms

Form	Estimate
ICC (1, 1)	.17
ICC (2, 1)	.29
ICC (3, 1)	.71
ICC (1, 4)	.44
ICC (2, 4)	.62
ICC (3, 4)	.91

from Case 2 or Case 3 studies is more likely. A consequence of this mistake is the underestimation of the true correlation ρ. For the same set of data, $ICC(1, 1)$ will, on the average, give smaller values than $ICC(2, 1)$ or ICC (3,1).

To help the reader appreciate the differences among these coefficients and also among the two coefficients to be discussed later, we apply the various forms to an example. Table 2 gives four ratings on six targets, Table 3 shows the ANOVA table, and Table 4 gives the calculated correlation estimates for various cases.

Given the choice of the appropriate index, tests of the null hypothesis—that $\rho = 0$—can be made, and confidence intervals around the parameter can be computed. When using $ICC(1, 1)$, the test that ρ is different from zero is provided by calculating $F_o = BMS/WMS$ and testing it on $(n-1)$ and $n(k-1)$ degrees of freedom. A confidence interval for ρ can be computed as follows: Let $F_{1-p}(i, j)$ denote the $(1-p)\cdot 100$th percentile of the F distribution with i and j degrees of freedom, and define

$$F_U = F_o \cdot F_{1-\frac{1}{2}\alpha}[n(k-1), (n-1)] \quad (4)$$

and

$$F_L = F_o / F_{1-\frac{1}{2}\alpha}[(n-1), n(k-1)]. \quad (5)$$

Then

$$\frac{F_L - 1}{F_L + (k-1)} < \rho < \frac{F_U - 1}{F_U + (k-1)} \quad (6)$$

is a $(1-\alpha)\cdot 100\%$ confidence interval for ρ.

When $ICC(2, 1)$ is appropriate, the significance test is again an F test, using $F_o = BMS/EMS$ on $(n-1)$ and $(k-1)(n-1)$ degrees of freedom. The confidence interval for $ICC(2, 1)$ is more complicated than that for $ICC(1, 1)$, since the index is a function of three independent mean squares. Following Satterthwaite (1946), Fleiss and Shrout (1978) have derived an approximate confidence interval. Let

$$\nu = \frac{(k-1)(n-1)\{k\hat{\rho}F_J + n[1 + (k-1)\hat{\rho}] - k\hat{\rho}\}^2}{(n-1)k^2\hat{\rho}^2F_J^2 + \{n[1 + (k-1)\hat{\rho}] - k\hat{\rho}\}^2},$$

where $F_J = JMS/EMS$ and $\hat{\rho} = ICC(2, 1)$. If we define $F^* = F_{1-\frac{1}{2}\alpha}[(n-1), \nu]$ and $F_* = F_{1-\frac{1}{2}\alpha}[\nu, (n-1)]$, then

$$\frac{n(BMS - F^*EMS)}{F^*[kJMS + (kn-k-n)EMS] + nBMS} < \rho < \frac{n(F_*BMS - EMS)}{kJMS + (kn-k-n)EMS + nF_*BMS} \quad (7)$$

gives an approximate $(1-\alpha)\cdot 100\%$ confidence interval around ρ.

Finally, when appropriate, $ICC(3, 1)$ is tested with $F_o = BMS/EMS$ on $(n-1)$ and $(n-1)(k-1)$ degrees of freedom. If we define

$$F_L = F_o / F_{1-\frac{1}{2}\alpha}[(n-1), (n-1)(k-1)]; \quad (8)$$

$$F_U = F_o \cdot F_{1-\frac{1}{2}\alpha}[(n-1)(k-1), (n-1)], \quad (9)$$

then

$$\frac{F_L - 1}{F_L + (k-1)} < \rho < \frac{F_U - 1}{F_U + (k-1)}$$

is a $(1-\alpha)\cdot 100\%$ confidence interval for ρ.

Decision 2: Can Effects Due to Judges Be Ignored in the Reliability Index?

In the previous section we stressed the importance of distinguishing Case 1 from Cases 2 and 3. In this section we discuss the choice between Cases 2 and 3. Most simply the choice is whether the raters are considered random effects (Case 2) or fixed effects (Case 3). Thus, under Case 2 we wish to generalize to other raters within some population, whereas under Case 3 we are interested only in a single rater or a fixed set of k raters. Of course, once the appropriate case is identified,

the choice of indices is between $ICC(2, 1)$ and $ICC(3, 1)$, as discussed before.

Most often, investigators would like to say that their rating scale can be effectively used by a variety of judges (Case 2), but there are some instances in which Case 3 is appropriate. Suppose that the reliability study (the G study) precedes a substantive study (the decision study in Cronbach et al.'s terms) in which each of the k judges is responsible for rating his or her own separate random sample of targets. If all the data in the final study are to be combined for analysis, the judges' effects will contribute to the variability of the ratings, and the random model with its associated $ICC(2, 1)$ is appropriate. If, on the other hand, each judge's ratings are analyzed separately, and the separate results pooled, then interjudge variability will not have any effect on the final results, and the model of fixed judge effects with its associate $ICC(3, 1)$ is appropriate.

Suppose that the substantive study involves a correlation between some reliable variable available for each target and the variable derived from the judges' ratings. One may either determine the correlation for the entire study sample or determine it separately for each judge's subsample and then pool the correlations using Fisher's z transformation. The variability of the judges' effects must be taken into account in the former case, but can be ignored in the latter.

Another example is a comparative study in which each judge rates a sample of targets from each of several groups. One may either compare the groups by combining the data from the k judges (in which case the component of variance due to judges contributes to variability, and the random effects model holds) or compare the groups separately for each judge and then pool the differences (in which case differences between the judges' mean levels of rating do not contribute to variability, and the model of fixed judge effects holds).

When the judge variance is ignored, the correlation index can be interpreted in terms of rater consistency rather than rater agreement. Researchers of the rating process may choose between $ICC(3, 1)$ and $ICC(2, 1)$ on the basis of which of these concepts they wish to measure. If, for example, two judges are used to rate the same n targets, the consistency of the two ratings is measured by $ICC(3, 1)$, treating the judges as fixed effects. To measure the agreement of these judges, $ICC(2, 1)$ is used, and the judges are considered random effects; in this instance the question being asked is whether the judges are interchangeable.

Bartko (1976) advised that consistency is never an appropriate reliability concept for raters; he preferred to limit the meaning of rater reliability to agreement. Algina (1978) objected to Bartko's restriction, pointing out that generalizability theory encompasses the case of raters as fixed effects. Without directly addressing Algina's criticisms, Bartko (1978) reiterated his earlier position. The following example illustrates that Bartko's blanket restriction is not only unwarranted but can also be misleading.

Consider a correlation study in which one judge does all the ratings or one set of judges does all the ratings and their mean is taken. In these cases, judges are appropriately considered fixed effects. If the investigator is interested in how much the correlations might be attenuated by lack of reliability in the ratings, the proper reliability index is $ICC(3, 1)$, since the correlations are not affected by judge mean differences in this case. In most cases the use of $ICC(2, 1)$ will result in a lower value than when $ICC(3, 1)$ is used. This relationship is illustrated in Tables 2, 3, and 4.

Although we have discussed the justification of using $ICC(3, 1)$ with reference to the final analysis of a substantive study, in many cases the final analytic strategy may rest on the reliability study itself. Consider, for example, the case discussed above in which each judge rates a different subsample of targets. In this instance the investigator can either calculate correlations across the total sample or calculate them within subsamples and pool them. If the reliability study indicates a large discrepancy between $ICC(2, 1)$ and $ICC(3, 1)$, the investigator may be forced to consider the latter analytic strategy, even though it involves a loss of degrees of freedom and a loss of computational simplicity.

Decision 3: What Is the Unit of Reliability?

The *ICC* indices discussed so far give the expected reliability of a single judge's ratings. In the substantive study (D study), often it is not the individual ratings that are used, but rather the mean of m ratings, where m need not be equal to k, the number of judges in the reliability study (G study). In such a case the reliability of the mean rating is of interest; this reliability will always be greater in magnitude than the reliability of the individual ratings, provided the latter is positive (cf. Lord & Novick, 1968).

Only occasionally is the choice of a mean rating as the unit of analysis based on substantive grounds. An example of a substantive choice is the investigation of the decisions (ratings) of a team of physicians, as they are found in a hospital setting. More typically, an investigator decides to use a mean as a unit of analysis because the individual rating is too unreliable. In this case, the number of observations (say, m) used to form the mean should be determined by a reliability study in pilot research, for example, as follows. Given the lower bound, ρ_L, on ρ from Inequality 6 or Inequality 7, whichever is appropriate, and given a value, say ρ^*, for the minimum acceptable value for the reliability coefficient (e.g., $\rho^* = .75$ or $.80$), it is possible to determine m as the smallest integer greater than or equal to

$$m = \frac{\rho^*(1 - \rho_L)}{\rho_L(1 - \rho^*)}.$$

Once m is determined, either by a reliability study or by a choice made on substantive grounds, the reliability of the ratings averaged over m judges can be estimated using the Spearman–Brown formula and the appropriate *ICC* index described earlier. When data from m judges are actually collected (e.g., in the D study following the G study used to determine m), they can be used to estimate the reliabilities of the mean ratings in one step, using the formulas below. In these applications, $k = m$. The formulas correspond to $ICC(1, 1)$, $ICC(2, 1)$, and $ICC(3, 1)$, and the significance test for each is the same as for their corresponding single-rater reliability index.

The index corresponding to $ICC(1, 1)$ is $ICC(1, k) = (BMS - WMS)/BMS$. Letting F_L and F_U be defined as in Equations 4 and 5,

$$1 - \frac{1}{F_L} < \rho < 1 - \frac{1}{F_U}$$

is a $(1 - \alpha) \cdot 100\%$ confidence interval for the population value of this intraclass correlation.

The index corresponding to $ICC(2, 1)$ is

$$ICC(2, k) = \frac{BMS - EMS}{BMS + (JMS - EMS)/n}.$$

The confidence interval for this index is most easily obtained by using the confidence bounds obtained for $ICC(2, 1)$ in the Spearman–Brown formula. For example, the lower bound for $ICC(2, k)$ is

$$\rho_L = \frac{k\rho_L^{**}}{1 + (k - 1)\rho_L^{**}},$$

where ρ_L^{**} is the lower bound obtained for $ICC(2, 1)$.

For $ICC(3, 1)$, the index of consistency for the mixed model case, the generalization from a single rating to a mean rating reliability is not quite as straightforward. Although the covariance between two ratings is $\sigma_T^2 - \sigma_I^2/(k - 1)$, the covariance between two means based on k judges is σ_T^2. As we pointed out before, under Case 3 no estimator exists for this term.

If, however, the Judge × Target interaction can be assumed to be absent, then the appropriate index is

$$ICC(3, k) = (BMS - EMS)/BMS.$$

Letting F_L and F_U be defined as in Equations 8 and 9,

$$1 - \frac{1}{F_L} < \rho < 1 - \frac{1}{F_U}$$

is a $(1 - \alpha) \cdot 100\%$ confidence interval for the population value of this intraclass correlation. $ICC(3, k)$ is equivalent to Cronbach's (1951) alpha; when the ratings of observers are dichotomous, it is equivalent to the Kuder–Richardson (1937) Formula 20.

Sometimes the choice of a unit of analysis causes a conflict between reliability considerations and substantive interpretations. A mean of k ratings might be needed for reliability, but the generalization of interest might be individuals.

For example, Bayes (1972) desired to relate ratings of interpersonal warmth to nonverbal communication variables. She reported the reliability of the warmth ratings based on the judgments of 30 observers on 15 targets. Because the rating variable that she related to the other variables was the mean rating over all 30 observers, she correctly reported the reliability of the mean ratings. With this index, she found that her mean ratings were reliable to .90. When she interpreted her findings, however, she generalized to single observers, not to other groups of 30 observers. This generalization may be problematic, since the reliability of the individual ratings was less than .30—a value the investigator did not report. In such a situation in which the unit of analysis is not the same as the unit generalized to, it is a good idea to report the reliabilities of both units.

Conclusion

It is important to assess the reliability of judgments made by observers in order to know the extent that measurements are measuring anything. Unreliable measurements cannot be expected to relate to any other variables, and their use in analyses frequently violates statistical assumptions. Intraclass correlation coefficients provide measures of reliability, but many forms exist and each is appropriate only in limited circumstances.

This article has discussed six forms of the intraclass correlation and guidelines for choosing among them. Important issues in the choice of an appropriate index include whether the ANOVA design should be one way or two way, whether raters are considered fixed or random effects, and whether the unit of analysis is a single rater or the mean of several raters. The discussion has been limited to a relatively pure data analysis case, k observers rating n targets with no missing data (i.e.. each of the n targets is rated by exactly k observers). Although we have implicitly limited the discussion to continuous rating scales, Feldt (1965) has reported that for $ICC(3, k)$ at least, the use of dichotomous dummy variables gives acceptable results. Readers interested in agreement indices for discrete data, however, should consult the Fleiss (1975) review of a dozen coefficients or the detailed review of coefficient kappa by Hubert (1977).

References

Algina, J. Comment on Bartko's "On various intraclass correlation reliability coefficients." *Psychological Bulletin*, 1978, *85*, 135–138.

Bartko, J. J. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 1966, *19*, 3–11.

Bartko, J. J. On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 1976, *83*, 762–765.

Bartko, J. J. Reply to Algina. *Psychological Bulletin*, 1978, *85*, 139–140.

Bayes, M. A. Behavioral cues of interpersonal warmth. *Journal of Consulting and Clinical Psychology*, 1972, *39*, 333–339.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements.* New York: Wiley, 1972.

Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, *16*, 407–424.

Feldt, L. S. The approximate sampling distribution of Kuder–Richardson reliability coefficient twenty. *Psychometrika*, 1965, *30*, 357–370.

Fleiss, J. L. Measuring the agreement between two raters on the presence or absence of a trait. *Biometrics*, 1975, *31*, 651–659.

Fleiss, J. L., & Shrout, P. E. Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika*, 1978, *43*, 259–262.

Haggard, E. A. *Intraclass correlation and the analysis of variance.* New York: Dryden Press, 1958.

Hayes, W. L. *Statistics for the social sciences.* New York: Holt, Rinehart & Winston, 1973.

Hubert, L. Kappa revisited. *Psychological Bulletin*, 1977, *84*, 289–297.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*, 151–160.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison–Wesley 1968.

Olkin, I., & Pratt, J. W. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 1958, *29*, 201–211.

Rajaratnam, N. Reliability formulas for independent decision data when reliability data are matched. *Psychometrika*, 1960, *25*, 261–271.

Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics*, 1946, *2*, 110–114.

Scheffé, H. *The analysis of variance.* New York: Wiley, 1959.

Sitgreaves, R. Review of *Intraclass correlation and the analysis of variance* by E. A. Haggard. *Journal of the American Statistical Association*, 1960, *55*, 384–385.

Snedecor, G. W., & Cochran, W. G. *Statistical methods* (6th ed.). Ames, Iowa: State University Press, 1967.

Winer, B. J. *Statistical principles in experimental Design* (2nd ed.). New York: McGraw-Hill, 1971.

Received January 9, 1978 ■

CONTENTS (continued)

As of February 1, 1978, Gene V Glass of the University of Colorado succeeded David A. Kenny as Associate Editor for methodological and statistical papers. As soon as most of the methodological articles in an issue have been reviewed by Glass, his name will replace Kenny's on the masthead.

3/12/79

VOLUME 86 / NUMBER 3 MAY 1979

Psychological Bulletin

(Continued on inside back cover)

LISHED BIMONTHLY BY THE AMERICAN PSYCHOLOGICAL ASSOCIATION
ISSN 0033-2909

The *Psychological Bulletin* publishes evaluative reviews and interpretations of substantive and methodological issues in the psychological research literature. The Journal reports original research only when it illustrates some methodological problem or issue. Discussions of methodological issues should be aimed at the solution of some particular research problem on psychology, but should be of sufficient breadth to interest a wide readership among psychologists; articles of a more specialized nature can be directed to the various statistical, psychometric, and methodological journals. The *Bulletin* does not publish original theoretical articles; these should be submitted to the *Psychological Review*.

Abstracts: All articles must be preceded by an abstract of 100–175 words. Detailed instructions for preparation of abstracts appear in the *Publication Manual of the American Psychological Association* (2nd ed.), or they may be obtained from the Editor or from APA Central Office.

Blind review: Because reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the author's identity or affiliation should be on separate pages. Every effort should be made to see that the manuscript itself contains no clues to the author's identity.

Manuscripts: Submit manuscripts in triplicate to the Editor, R. J. Herrnstein, *Psychological Bulletin,* Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, according to instructions provided below.

Instructions to Authors: Authors should follow the directions given in the *Publication Manual of the American Psychological Association* (2nd ed.). Instructions on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors are requested to refer to the "Guidelines for Nonsexist Language in APA Journals" (Publication Manual Change Sheet 2, *American Psychologist,* June 1977, pp. 487–494) before submitting manuscripts to this journal. All manuscripts should be submitted in duplicate and both copies should be clear, readable, and on paper of good quality. Dittoed copies are not acceptable and will not be considered. Authors are cautioned to carefully check the typing of the final copy and to retain a copy of the manuscript to guard against loss in the mail.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1979: $40 domestic, $42 foreign, $7 single issue. APA member rate: $15. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Dept., APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in January) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Printed in the U.S.A. Second-class postage paid at Arlington, Va., and at additional mailing offices.

Vol. 86, No. 3 May 1979

Psychological Bulletin

Equity Theory and the Cognitive Ability of Children

J. G. Hook and Thomas D. Cook
Northwestern University

A number of studies indicate that preadolescents allocate more rewards to those who have done more work. Adams's equity theory is most often used to explain this finding. One assumption of equity theory is that persons compute ratios and compare them for proportionality. However, research on logico-mathematical development indicates that children do not solve problems of proportionality until they are 11–15 years old. This suggests that equity theory may not be an adequate explanation of how children allocate rewards in experiments on equity. Children's allocation behaviors do change with age, from the possibly self-interested or equal allocations of children under 6 years, to the descriptive ordinal equity allocations of 6- to 12-year-olds, to the possibly proportional allocations of persons 13 years and older. This sequence is consistent with the normal sequence of logico-mathematical development, suggesting that observed allocation behaviors may be a function of cognitive ability as well as manipulated situational variables.

Adams's (1965) equity theory is a cognitive social comparison theory. It is concerned with the cognitive activity of the individual who is confronted with a problem of distributive justice. We call this individual the *allocator*. Adams assumed that the allocator constantly compares persons, including the allocator, on two dimensions: outcomes and inputs. Inputs are contributions (e.g., labor) or attributes (e.g., being well educated) that justify claims on outcomes. Outcomes are rewards or desirable things, which can be either tangible (e.g., pay) or intangible (e.g., love). Adams assumed that the allocator first forms ratios of outcomes to inputs for each comparison person. Next, the allocator compares the ratios. If they are equivalent, or proportional, equity obtains. If they are not proportional for any two comparison persons, inequity is assumed to result. Thus, inequity obtains when

$$\frac{\textit{Outcomes for Person X}}{\textit{Inputs by Person X}} \neq \frac{\textit{Outcomes for Person Y}}{\textit{Inputs by Person Y}}.$$

If inequity obtains, Adams assumed that the allocator will recognize it and feel a sort of discomfort. This discomfort is then assumed to motivate the allocator to restore equity, either directly (e.g., by redistributing rewards) or by cognitive distortion.

Adams's equity theory incorporates the value principle that there should be a rela-

The first author was partially supported by the Law and Social Sciences Program of Northwestern University.

Requests for reprints should be sent to Thomas D. Cook, Department of Psychology, Northwestern University, Evanston, Illinois 60201.

tionship of proportionality between inputs and outcomes, between work and reward; that is, rewards should be allocated according to merit. Each allocator, however, has substantial latitude in determining the content of equity, because the allocator alone determines which particular factors to include as inputs and outcomes and how each should be weighted. Adams's theory is not an ethical theory because he did not say equitable allocations were good. Rather, he predicted that people would behave as if they believed equitable allocations were good. According to Adams, people behave that way because they are responsive to a social norm that prescribes equitable allocations and proscribes other allocations. The equity norm is presumably the cause of the allocator's discomfort when the allocator recognizes inequity.

Few equity theorists believe that the equity norm is always applicable. Rather, the norm is presumed to guide behavior in certain situations and not in others. When the equity norm is not situationally appropriate, other norms may be relevant. The equality norm (Sampson, 1969), for example, commands that persons receive the same outcomes, no matter what their inputs. Self-interest, which is perhaps a genetically influenced norm (Campbell, 1972), commands that persons maximize their own share. A norm for altruism (Bryan, 1972) commands the opposite. Other hypothesized norms, less relevant to this article, include reciprocity (Schopler, 1970; Staub, 1972) and "to each according to his need" (Berkowitz, 1972; Leventhal, Weiss, & Buttrick, 1973).

Most allocation research has been designed to identify the personal and situational variables that influence an allocator's choice of norms. For example, research indicates that under some conditions females make equal allocations of reward regardless of relative work (Uesugi & Vinacke, 1963), whereas males make equitable allocations (Vinacke & Stanley, Note 1). Under different conditions the reverse can be true (Kidder, Bellettirie, & Cohn, 1977). Other empirical work suggests that allocators make equal allocations when group harmony is a salient concern (Smith & Cook, 1973), but make equitable allocations when productivity is a salient concern (Leventhal, Michaels, & Sanford, 1972).

In the first study in which equity theory was used to predict children's allocation behaviors, Leventhal and Anderson (1970) led individual children to believe that they were members of dyads doing work for the experimenters. One independent variable was the sex of the child. The second involved a manipulation of the amount of work a child did relative to the amount he was led to believe that his (unseen) partner had performed. This relative-work information was the only information each child was given about his partner. Each child was assigned to one of three work conditions: superior work (three times as much work as the partner in a 15:5 ratio), equal work (same amount of work in a 15:15 or 5:5 ratio), and inferior work (one third as much work in a 5:15 ratio). The dependent measure was the number of rewards, out of a total of 20, that each child kept for himself when he was asked to distribute the rewards anonymously between himself and the other child in any way he desired. Equity theory was used to predict that the children would keep the same proportion of rewards as the proportion of work they had done. For example, the inferior-work child's *equity cognition* appeared first as the dilemma

$$\frac{\textit{My work} = 5}{\textit{My reward} = ?} = \frac{\textit{Other's work} = 15}{\textit{Other's reward} = ?},$$

to be resolved as $5/5 = 15/15$. Thus, equity theory, which assumes that the allocator forms ratios in his head and compares them for *proportional equity*, predicts that superior-, equal-, and inferior-work subjects should keep 15, 10, and 5 rewards, respectively.

Table 1 shows the results of the Leventhal and Anderson study. Recall that the scores could range from 0 to 20, with 10 indicating that the child divided rewards equally between himself and the other child. The means for females do not differ significantly from each other or from 10. This suggests that under the conditions of this experiment the

female allocators were influenced by an equality norm. Males in the superior-work (15 units) condition kept significantly more for themselves than did males in the two equal-work conditions. This result seems to support an equity norm interpretation. However, it does not support an equity interpretation in the strictest sense of Adams's theory. Rather, it is an *ordinal equity,* in which rank order is preserved from the work dimension to the reward dimension. The mean value of 12.7 is halfway between the proportional equity predicted by equity theory (i.e., 15) and the *equality norm* (10). We henceforth assume that such behavior reflects ordinal equity.

It is crucial to understand that even if the superior workers kept larger rewards and the inferior workers smaller rewards (although the latter did not occur in this study), such results would not necessarily be consistent with equity theory. This is because the theory requires that workers keep the same proportion of rewards as their proportion of work. Thus, results like those found with the superior-work males of Leventhal and Anderson cannot readily be interpreted in terms of equity theory alone. One would have to assume that proportional equity is operating together with other forces influencing allocation, or that some alternative (e.g., ordinal equity) is involved.

Table 1
Mean Number of Rewards Kept by Subjects in Leventhal and Anderson (1970)

Subject's work unit	Partner's work unit: 5	Partner's work unit: 15
Females		
5		
Theoretical no. rewards	10	5
Actual *M* no. rewards	10.7	10.1
15		
Theoretical no. rewards	15	10
Actual *M* no. rewards	10.3	10.2
Males		
5		
Theoretical no. rewards	10	5
Actual *M* no. rewards	10.6	11.1
15		
Theoretical no. rewards	15	10
Actual *M* no. rewards	12.7	10.3

Before we move to the literature review it is important that we briefly review two key aspects of equity theory. In the Leventhal and Anderson (1970) study, equity was inferred from the reward allocations of children from different input groups, and there was no explicit comparison with the precise scores predicted by the proportional equity theory. In other words, the hypotheses and independent variables were derived from a theory that assumes interval scales for the input and outcome variables, but inferences about equity were made merely on the basis of ordinal group differences on the dependent variable. It is as though equity researchers are content to infer equity if the mean differences are in the expected direction, irrespective of whether they differ from the theoretically specified numerical differences. The justifications for such slippage are presumably (a) that experiments cannot control all inputs and outcomes and that some of the nonmanipulated variables that subjects make relevant to a particular equity ratio may codetermine their allocative behavior; (b) that equity may be aroused simultaneously with other norms that codetermine allocation; and (c) that equity theory should be understood as a general metaphor for studying allocative behavior rather than as a formal theory pretending to completeness and specificity. Unfortunately, equity theorists are rarely explicit about the reasons for inferring equity from ordinal patterns of data that are in line with, but different from, the more exact interval-scale predictions that the theory appears to make.

Second, equity theory makes somewhat extreme assumptions about the cognitive activity of the allocator. Recall that the theory states that an allocator feels discomfort when equity ratios are not proportional. Almost all tests of equity leave this discomfort unmeasured and thus assumed. Logically prior to this feeling is the assumption that the allocator sets up and compares ratios. Still prior is the assumption that the allocator is competent to set up and compare ratios. Then, of course, there is the large assumption that the allocator is able to conceptualize heterogeneous outcomes (e.g., satisfaction or dol-

Table 2
Allocation Literature

Study	Age in years	Class	Result	Equity quotient[a]
Masters (1968)	3–5	P	Self-interest or equality	
Nelson & Dweck (1977)	4	P	Equality	18
Peterson, Peterson, & McDonald (1975)	4	3rd	Equality	11
				16
Peterson, Peterson, & McDonald (1975)	4	3rd	Equality	14
				24
Antone & Hendricks (Note 3)	3–7	P	Self-interest (3–6)	
			Equality (7)	
Lane & Coon (1972)	4	P	Self-interest	4
	5	P	Equality	6
Handlon & Gross (1959)	4–5	P	Self-interest	
Lerner (1974)	5	P	Equality	9[b]
				19[c]
Lerner (1974)	5	3rd	Ordinal equity	40[b]
			Equality	22[c]
Hook (1978)	5	P	Self-interest	22
Leventhal & Anderson (1970)	5	P	Equality	16[b]
				1[c]
Leventhal, Popp, & Sawyer (1973)	5	3rd	Equality	20[b]
			Ordinal equity	40[c]
Leventhal, Popp, & Sawyer (1973)	5	3rd	Ordinal equity	52[b]
				36[c]
Coon, Lane, & Lichtman (1974)	5	3rd	Ordinal equity	46
Larsen & Kellogg (1974)	4–8	P	Equality	
Olejnik (1976)	5	3rd	Ordinal equity	53
Lerner (1974)	6	P	Ordinal equity	40[b]
			Equality	13[c]
Olejnik (1976)	6	3rd	Ordinal equity	52
Libby & Garrett (1974)	6	P	Ordinal equity	60
Streater & Chertkoff (1976)	6	P	Ordinal equity	
Streater & Chertkoff (1976)	6	3rd	Ordinal equity	
Anderson & Butzin (1978)	6	3rd	Ordinal equity	
Hook, Brockett, & Smith (Note 4)	6–7	3rd	Ordinal equity	20
Olejnik (1976)	7	3rd	Ordinal equity	54
Coon, Lane, & Lichtman (1974)	7	3rd	Ordinal equity	64
Streater & Chertkoff (1976)	8	3rd	Ordinal equity	
Streater & Chertkoff (1976)	8	3rd	Ordinal equity	
Leventhal, Popp, & Sawyer (1973)	8	3rd	Ordinal equity	47[b]
				27[c]
Leventhal, Popp, & Sawyer (1973)	8	3rd	Ordinal equity	27[b]
				27[c]
Olejnik (1976)	8	3rd	Ordinal equity	66
Tompkins & Olejnik (1978)	7–9	3rd	Ordinal equity	31[d]
			Equality	20[e]
Anderson & Butzin (1978)	8	3rd	Ordinal equity	
Cohen & Sampson (Note 5)	3–12	3rd		
Hook, Brockett, & Smith (Note 4)	8–9	3rd	Ordinal equity	42
Coon, Lane, & Lichtman (1974)	9	3rd	Ordinal equity	68
Hook (1978)	9	P	Ordinal equity	56
Handlon & Gross (1959)	9–11	P	Equality	
Libby & Garrett (1974)	10	P	Ordinal equity	53
Lerner (1974)	10	P	Ordinal equity	50[b]
				50[c]
Lerner (1974)	10	3rd	Ordinal equity	60[b]
				50[c]
Anderson & Butzin (1978)	10	3rd	Ordinal equity	
Benton (1971)	1–12	P	Ordinal equity and equality	

Table 2 (*continued*)

Study	Age in years	Class	Result	Equity quotient[a]
Hook, Brockett, & Smith (Note 4)	10–11	3rd	Ordinal equity	48
Morgan & Sawyer (1967)	10–12	P	Ordinal equity	58
Coon, Lane, & Lichtman (1974)	11	3rd	Ordinal equity	68
Streater & Chertkoff (1976)	12	3rd	Ordinal equity	
Hook, Brockett, & Smith (Note 4)	12–13	3rd	Ordinal equity	68
Hook (1978)	13	P	Proportional equity	94
Garrett & Libby (1973)	14	P	Proportional equity	96
Anderson (1976)	Adult	3rd	Proportional equity	
Leventhal & Michaels (1969)	Adult	P	Ordinal equity	43
				40
Leventhal, Weiss, & Long (1969)	Adult	P	Proportional equity	109
Leventhal & Lane (1970)	Adult	P	Proportional equity	79[b]
				79[c]
Lane, Messe, & Phillips (1971)	Adult	P	Proportional equity	
Kahn (1972)	Adult	P	Ordinal equity	61[b]
				48[c]
Cohen (1974)	Adult	P	Proportional equity	183
Shapiro (1975)	Adult	P	Proportional equity	80
Von Grumbkow, Deen, Steensma, & Wilke (1976)	Adult	3rd	Proportional equity	78
Reis & Gruzen (1976)	Adult	P	Ordinal equity	48
Kidder, Bellettirie, & Cohn (1977)	Adult	P	Ordinal equity	

Note. P = studies in which subjects asked to allocate rewards were also participants in the work and potential reward recipients; 3rd = studies in which subjects asked to allocate rewards were third party observers of others' work.
[a] Decimal points omitted.
[b] Subjects were male.
[c] Subjects were female.
[d] Correspondent problem.
[e] Ratio problem.

lars) and heterogeneous inputs (the quantity and quality of work or qualifications) and is able to weld them into a single, unidimensional scale of outcomes or inputs. It may or may not be true that adults go through such a cognitive process, of which they may well be capable under some conditions This article is concerned with whether children are capable of such cognitive activities and whether they actually perform them.

Allocation Literature: Children and Adults

Framework for the Literature Review

Table 2 is a summary of the allocation literature. Two types of studies are included. The Leventhal and Anderson (1970) study is an example of one type, in which subjects are allowed to compare the work inputs of persons and then are asked to distribute rewards (outcomes) among the persons who have performed work. The Libby and Garrett (1974) study is an example of the second type, in which the allocator is asked to distribute rewards between persons who did equal work throughout but were unequally rewarded in a previous allocation.

Two types of allocation studies are excluded because they do not involve work comparisons and reward allocations among the same persons. First, in "helping" studies, allocators are asked to donate some of their rewards to a third party who did not work (Long & Lerner, 1974; Miller & Smith, 1977; Staub, Note 2) or are asked to share an unearned gift with a friend or stranger (Ugurel-Semin, 1952; Wright, 1942a, 1942b). Second, in some equity studies, all the work and outcome values are told to subjects, making the computation of an equity formula a fait accompli. The dependent measures are

such factors as subjects' ratings of the allocator's attractiveness or fairness (Brickman & Bryan, 1975, 1976) or the quality of the subject's continued performance under some preestablished allocation rule (Lawler, 1968).

The explicit condition from Nelson and Dweck (1977) is excluded from Table 2. This is because children were instructed to give out rewards "so that you get the right amount for doing this much work" (p. 194), and then had a chance to physically copy in their reward allocation what the work differences had been. In the research reported in Table 2, children were instructed to allocate rewards in a fair manner or in whatever manner they wanted, and copying was not explicitly requested.

The studies in Table 2 are arranged according to the ages of the subjects. Studies of multiple age groups are reported multiple times in appropriate age positions. The column labeled Class indicates whether the subjects who were asked to allocate rewards were participants in the work and potential reward recipients (P studies) or were disinterested third party observers of others' work (3rd studies).

The column labeled Result classifies each study with respect to the allocation principle the mean scores seem to indicate the subjects followed. The four classes of results are self-interest, equality, ordinal equity, and proportional equity. Self-interest could only occur in P studies, which means the subjects kept more for themselves than they allocated to others, no matter how much work each had done. Equality, ordinal equity, and proportional equity are defined with respect to the Equity Quotient column. The equity quotient is a numerical index of the extent to which the study results are consistent with Adams's proportional equity. The quotient is the ratio of the percentage point differences between experimental conditions predicted by proportional equity theory to the differences between experimental conditions in the observed allocation behaviors. For example, in Leventhal and Anderson (1970) the inferior-work males should have kept 25% of the reward because they did 25% of the work. The superior-work males should have kept 75% of the reward because they did 75% of the work. Thus, the gap between the two groups predicted by proportional equity theory is 50 percentage points (75 − 25 = 50). The actual difference between the two experimental conditions, in terms of their actual allocation behaviors, was 8 percentage points. This is because the interior-work group kept an average of 55.5% of the reward, whereas the superior-work group kept an average of 63.5%. The ratio of the observed gap (8) to the predicted gap (50) is therefore .16 (8/50), which is the quotient reported in Table 2 for the males from Leventhal and Anderson. (Actually, all values in the table have been multiplied by 100 to avoid decimal points.)

The ratio in Leventhal and Anderson can be expressed visually as

Observed: 55.5.....63.5 = 8

Predicted: 25.0.............75.0 = 50.

A quotient of zero means that subjects did not differentiate their reward allocations:

Observed: 50–50 = 0

Predicted: 25.0.............75.0 = 50.

A quotient of 100 means that subjects made reward allocations precisely as predicted by proportional equity theory:

Observed: 25.............75 = 50

Predicted: 25.............75 = 50.

A quotient of 50 is the prototypical ordinal equity relationship, with observed scores falling between the equality and the proportional equity predictions:

Observed: 37.5.....62.5 = 25

Predicted: 25.............75 = 50.

In Table 2, then, scores between 0 and 25 are labeled *equality* or *self-interest*. Scores between 26 and 75 are called *ordinal equity*, and scores greater than 75 are labeled *proportional equity*.[1]

[1] If a study included only one work condition (Larsen & Kellogg, 1974) or more than two persons in the reward allocations (Streater & Chertkoff,

Relationship Between Age and Equity

The studies in Table 2 and other allocation studies suggest that a number of independent variables influence children's reward allocations. Children's allocation behaviors may be influenced by whether the allocator is a participant or a third party, whether the participants are in a team or nonteam relationship (Lerner, 1974), and whether they are cooperating or competing (Crockenberg, Bryant, & Wilce, 1976), expect future interactions with each other (Dreman, 1976), are male or female (Leventhal & Anderson, 1970), believe the experimenter will or will not evaluate their allocations (Leventhal, Popp, & Sawyer, 1973), and have insufficient, sufficient, or oversufficient total reward to allocate (Coon, Lane, & Lichtman, 1974).

However, the independent variable of special interest for this article is age, especially as it mediates logical development. The most striking age-linked feature of Table 2 is that the proportional equity responses predicted by equity theory are entirely absent in studies of subjects under 13 years of age. On the other hand, the results provide strong support for an alternative, ordinal equity interpretation with children. Three sets of studies are especially interesting in this regard because the same investigators employed the same designs with children 12 years of age or under and with children or adults of 13 years and over. The equity quotient scores for 5.9-year-old children in Leventhal and Anderson (1970) were 16 for males and 1 for females, whereas adult quotients in Leventhal and Lane (1970) were 79 for both males and females. Libby and Garrett's (1974) 6- and 8-year-olds gave quotients of 60 and 53, respectively, whereas Garrett and Libby's (1973) 14-year-olds gave a quotient of 96. Hook's (1978) 5-, 9-, and 13-year-olds gave quotients of 22, 56, and 94, respectively. The average equity quotient in studies of children under 13 years was 36.9, whereas the average quotient for 13-year-olds and above was 79.9. Proportional equity, therefore, appears to be restricted to the teen years and beyond.

How can one explain the ordinal equity of childhood? Five possibilities are discussed below.

Confounded inputs interpretation. One interpretation is that children use other inputs in addition to work in their proportional equity equations. For example, suppose most children felt that need should be weighed as heavily as work. Since most studies provided no information about the other's need, the average subject might assume that he and the other were equal on need. If so, the superior-work subject would have total inputs somewhat greater than the other, but not so much greater than if he had used only work in forming inputs. Thus, in allocating rewards proportional to inputs, the superior workers would make what, to the researcher who was not aware of the hidden need input, looked like an ordinal equity allocation. Yet the allocation might be based on proportional equity. (As with inputs, a subject could also add hidden outcomes or rewards to those manipulated by an experimenter.) The uncontrolled-inputs-and-outcomes problem is a major weakness of equity theory, making falsification extremely difficult. No matter what reward allocation a subject made, it could be argued that he was behaving equitably on the basis of some uncontrolled intruding input or outcome.

Nonisomorphism of physical and psychological scales. A second interpretation follows from the possibility that preadolescents transform the physical work scales into psychological scales of a different form. If so, one can imagine circumstances in which descriptive ordinal allocation data can be interpreted as consequences of proportional thinking of the type suggested by Adams (1965). But it is proportional thinking based on psychological scales that are not linear functions of the corresponding physical scales.

Intra-allocator norm combination. A third interpretation of the ordinal equity data of

---

1976), no quotient could be calculated. Also, in two studies (Anderson, 1976; Anderson & Butzin, 1978) the independent manipulations were on an ordinal scale (e.g., one child gave little effort and the other average effort) rather than on a ratio scale (e.g., 15/5), which did not allow quotient calculation. These few studies are classified on logical grounds.

younger children is that they are simultaneously aware of the allocations appropriate to equality and proportional equity norms but resolve any dilemmas associated with the different norms by taking the average. Such a compromise yields an ordinal equity allocation. This interpretation holds that older children have clear preferences for proportional equity over equality norms, which is why their data reflect proportional rather than ordinal equity. No data exist of which we are aware that probe whether children under about 13 years of age consider both equality and proportional equity norms before deciding on a compromise between the two.

Interallocator norm combination. A fourth interpretation is that some younger children make equality allocations, whereas others make proportional equity allocations. The average scores in this case would look like ordinal equity allocations. However, such an interpretation requires a bimodal distribution of allocations. Hook's (1978) data are not bimodally distributed, and no published study mentions a bimodal distribution or reports group differences in variances for the allocation measure. Consequently, the interpretation based on the interallocator norm combination does not seem likely.

Inability interpretation. A fifth interpretation, the primary concern of this article, is that children under the age of 13 are typically incapable of solving problems of ratio proportionality. If this were true, it would rule out the first three interpretations, since all three assume that half or more of the children are engaged in the cognitive, ratio-proportionality activity posited by equity theory. The inability interpretation implies that children do not weight several norms to decide which to apply in a given situation because it postulates that children are incapable of recognizing proportional equity. However, the inability explanation does not rule out the possibility that children under 13 years of age may be capable of acquiring and expressing work–reward relationships in an ordinal equity sense, assuming that ordinal equity is an earlier acquisition than is proportional equity.

To decide whether children under the age of 13 can solve problems of proportional or ordinal equity, we turn to the literature on logico-mathematical development: the interview work of Piaget's school and the more statistical, normative work of other investigators.

Development of Proportional Thought

Table 3 is a list of Piaget's studies of the development of logico-mathematical proportion. The column labeled Task refers to the particular problem used by Piaget to explore proportional thought. According to his theory, proportional thought is a cognitive structure that is central to, or embedded in, all kinds of mathematical, physical, and social problems. Piaget has not studied proportional thought directly. Rather, he has studied it in the context of his research on such diverse yet proportion-dependent areas as geometry, chance, time, and functions. Piaget and his colleagues usually conduct clinical interviews with children of different ages and classify their responses into age-related stages. Each stage is supposed to represent a qualitatively different approach, and the stages are thought to unfold in an invariant sequence, which culminates in mature proportionality, the stage at which all logical components have been integrated into a coherent structure. The column labeled Attainment Age in Years refers to the approximate age that Piaget's subjects attained the stage of proportional thought. Piaget was not very explicit about how the attainment age was defined, and it is obvious that the age depended on certain task factors, for example, perceptual and memory requirements. Therefore, the attainment ages should be considered estimates.

To illustrate Piaget's approach, we consider two of his studies. In studies of the development of the concept of speed, Piaget (1970) asked children to trace and time the movements of two objects, A and B, that ran in succession. The children were then asked which moved faster. The youngest children did not transfer the ordinal rank of the objects from distance traveled to speed. In other words, if A moved farther than B in

Table 3
Piaget's Research on the Development of Proportional Thought

Study	Task	Age studied in years	Attainment age in years
Piaget & Inhelder (1956)	Constructing similar triangles and rectangles	6–14	11–12
Piaget (1957)	Constructing similar triangles	6–14	11–12
Inhelder & Piaget (1958)	Guessing the size of shadows cast by rings	6–14	12
	Equilibrium creation on a balance arm with different weights	6–14	13
Piaget, Inhelder, & Szeminska (1960)	Building equal-volume block structures on unequal-area bases	6–14	11–12½
Piaget, Grize, Szeminska, & Vinh Bang (1968)	Making different-size fishes equally well fed	6–14	10–11
	Covarying different-size circles with different line positions	7–13	12–13
	Relative movements of wheels of different diameters	6–13	11–13
	Relative movements of objects pulled by different-diameter pulleys	6–14	12–14
	Equilibrium on a balance arm	6–14	12–13
Piaget (1970)	Drawing lines to symbolize relative speeds of objects	5–14	12–13
Piaget (1974)	Stretching elastic bands and guessing relative lengths of segments	5–14	13–14
	Maintaining angles formed by strings held fast with different angles	5–14	13–14
	Guessing the number of small unit beakers necessary to fill larger beakers	5–14	13–14
Piaget & Inhelder (1975)	Probabilities of lottery drawing outcomes from beakers containing various ratios of elements	5–14	11–12

the same amount of time, these children did not necessarily say that A moved faster. The 7- to 9-year-olds did make the transformation, but only if either distance traveled or time was held constant. If time and distance were varied simultaneously, the 7- to 9-year-olds were unable to set up the necessary ratios of distance to time for comparison:

$$\frac{\textit{Distance for A}}{\textit{Time for A}} = \frac{\textit{Distance for B}}{\textit{Time for B}}.$$

This problem is of course highly analogous to the equity theory problem, if one substitutes outcomes for distance and inputs for time.

The 10- and 11-year-olds could often solve the two-variable problem for simple 2:1 ratios. However, their solutions may have been intuitive instead of formal. They did not seem to be able to explain why they guessed as they did. Only 12- to 14-year-olds could solve the problems for more complex ratios and state the principle that distance divided by time equals speed.

Piaget (1974) obtained analogous results by asking children to estimate the lengths of elastic band segments. The bands were divided into Segments A and B, distinguished by color. Segment A was longer than Segment B by various ratios. For example, in a 2:1 ratio problem, the elastic band at rest was 3 cm long, with segments of 2 cm (A) and 1 cm (B). The band was then stretched to lengths that were multiples of its rest length, for example, doubled to 6 cm. Before the children could examine the segments of the expanded band, they were asked how long each segment should be. The youngest children (4–5 years) failed to preserve ordinality through the transformation, not realizing that A should be longer than B. Somewhat older children (6–9 years) noted that A should be longer, but did not preserve the ratio relationships between A and B. Rather,

Table 4
Anglo-American Research on the Development of Proportional Thought

Study	Task	Age studied in years	Attainment age in years
Lovell (1961)	Equilibrium on a balance arm	11–15	13–15
	Guessing the shadows cast by rings	11–15	13–15
Lunzer (1965)	Number series and analogies	9–17	13–15
Lovell & Butterworth (1966)	Equilibrium on a balance arm		13–15
	Guessing the shadows cast by rings	11–15	13–15
	Number analogies	11–15	13–15
	Calculating the relation between polygan angles and the number of sides	11–15	13–15
	Calculating the areas of similar triangles	11–15	13–15
Bruner & Kenney (1966)	Filling similar and dissimilar beakers to same and different proportions full	5–7, 9, 11	11
Steffe & Parr (Note 6)	Pictorial and symbolic ratio forms		13–15
Fischbein, Pampu, & Manzat (1970)	Finding equal ratios of beads in two beakers	5–6 9–10 12–13	12–13
Brainerd & Allen (1971)	Density conversion	10–11	80% failed proportionality problem
Lee (1971)	Equilibrium on a balance arm	5–17	13
	Guessing shadows cast by rings	5–17	13
Tomlinson-Keasey & Keasey (1974)	Equilibrium on a balance arm	11–12 18–20	Only in the 18–20 group
Webb (1974)	Equilibrium on a balance arm	5–11 (All IQs > 160)	None
Chapman (1975)	Picking the container with higher proportions of certain-color beads	6 8 10 College	Only in college group

they preserved the additive relationship. If Segment A was 1 cm longer than Segment B before the stretching $(2 - 1 = 1)$, it should have been 1 cm longer afterwards $(4 - 3 = 1)$. The oldest children discovered the ratio solution first on an intuitive level and with small (2:1) ratios and then, at about 13 years of age, on a formal basis with complex ratios.

Piaget's other studies yielded results quite analogous to the two examples. The developmental stages seem to unfold in the same sequence no matter what task is used. The age of attainment varies somewhat depending on the task, but in all cases it is between 10 and 15 years of age, averaging about 13 years of age.

Piaget's methodology has been criticized as being too clinical. Table 4 is a summary of research carried out by non-Piagetian psychologists. Many of these studies employed the large samples, statistical analyses, and standardized problem presentation and response modes that are missing in the Piagetian research. The attainment ages mentioned in the table refer to the age at which half or more of the subjects had demonstrated proportional thought, if it was possible to glean this information from the published reports. In general, the studies in

Table 4 are consistent with Piaget's results. They document the same acquisition stages and the fact that mature proportional thought is not developed until adolescence. However, these attainment ages are a year or two later than those reported by Piaget.

Implications of the Literature Review

Our review of the literature on logic–mathematics development indicates that the average child under 13 years of age does not solve problems of ratio proportionality. There is no evidence with physical and mathematical problems that preteens cognitively construct and compare ratios, just as there is no evidence for similar cognitive activity in the allocation literature. Our review offers no support for the equity theory assumptions that preteens form cognitive ratios and feel discomfort in the absence of proportionality. The major methodological and theoretical implications of the review are now spelled out.

Methodological Implications

Our negative conclusions about the relevance of Adams's (1965) proportional equity theory for children under 13 years have to be tentative. They must be interpreted in the light of several issues. The research summarized in Tables 3 and 4 does not necessarily demonstrate that children under 13 years are incapable of proportional solutions to logico-mathematical problems. It may only show that they do not perform behavior commensurate with such thinking. The studies cited in these tables were open ended, and children were asked to respond in whatever manner they wished. Generally, they were not told that there was a right answer that they must find; no incentives were provided to induce children to perform in ways indicative of proportional thinking; and the studies did not follow the training or enrichment paradigm whereby a child is told a correct response and is reinforced for replicating it. Brainerd and Allen (1971) claimed that corrective feedback over a series of trials on conservation of density induced this capacity in 10- and 11-year-olds who at first failed to conserve and did not seem to have the capacity. The trained children not only conserved density at a higher rate than a control group but also apparently generalized their learning to problems in conservation of volume. Varying training, instructional sets, and incentives to reveal capacities that are not spontaneously manifested may provide evidence of proportional capacity in younger children. We do not know, since the appropriate studies have not yet been performed.[2] Until they are, it will not be clear whether the age norms for both allocative behaviors and logico-mathematical skills are due to the biological constraints of maturation or to environmental factors that lead children not to perform behavior of which they are indeed capable under certain conditions. (See Brainerd, 1978, for a discussion of this problem.)

Even if one assumes that preteens are incapable of proportional responses to logico-mathematical problems, does this mean they are incapable of such responses to allocation dilemmas? Our response must be, "We are not yet sure," largely because of the possibility of confounded inputs and nonisomorphic psychological and physical scales. But on the other hand, our review does suggest that behaviors in the allocative and logico-mathematical domains are highly congruent. Also, a number of studies (Damon, 1975; Hook, 1978; Tompkins & Olejnik, 1978) document correlations between responses to the two types of problem. However, such relationships do not show which type, if either, is prior. Future research is needed to test the idea, implicit in our review, that the logico-mathematical proportionality concept develops either prior to, or simultaneous with, the allocational proportionality concept.

Some experimental features may permit children to make allocations that are entirely consistent with proportional equity but are not the result of engaging in proportional cognitions. The first such feature in-

[2] The experiment by Brainerd and Allen (1971) did not involve feedback on actual numerical distributions, and their subjects could have given proportionality responses without ratio cognitions.

volves the recognition of equity as a dependent behavior rather than the creation of equity. Brickman and Bryan (1976), for example, showed children videotapes in which characters transferred rewards from one person to another and either created or eliminated proportional equity. The children then rated the transfer agent. Agents who created equity were viewed more favorably than those who eliminated it. Thus, the children did not create an equitable allocation, but recognized one.

A second feature involves the use of correspondent allocation problems. Using the Leventhal and Anderson (1970) paradigm, suppose that two persons contribute 15 and 5 inputs, respectively, and then must distribute 20 outcome rewards. A proportional equity allocation could be created by setting outcomes in exact correspondence to inputs for each person: 15 for 15 and 5 for 5. This allocation requires no knowledge of ratios. Nelson and Dweck (1977), for example, asked 4-year-olds who had done either nine or three units of work to divide 12 rewards. In the conditions in which they specifically instructed children to make allocations consistent with work, all of the children made proportional allocations. Although Nelson and Dweck argued that this demonstrated the capacity of even 4-year-olds to allocate in an equitable manner, Anderson and Butzin (1978) noted that such proportional allocations could have been due to correspondence procedures involving no social comparison or proportional cognition at all. Also, Nelson and Dweck observed proportional allocations only when they explicitly requested work–reward correspondence. In other conditions, Nelson and Dweck and other researchers using correspondence designs (Streater & Chertkoff, 1976; Tompkins & Olejnik, 1978) did not observe proportional allocations.

What if, however, two persons contributed three and one inputs, respectively, and were asked to allocate 20 rewards? Children could give 3 rewards to one worker and 1 to the other and then repeat the process until the 20 rewards were exhausted. This problem demands *iteration* but not ratio cognitions. It would be more plausible to assume ratio cognitions if the two workers contributed 6 and 2 units of work before allocating 20 rewards.

A third feature involves the *ordinal scaling* of the independent variable. Anderson and Butzin (1978) used adjectives to describe the work efforts of two persons. Persons X and Y were said to have tried "a little bit," "kind of hard," and "very hard." Children aged 6-, 8-, and 10-years-old were asked to allocate 20 candy rewards to X and Y under all nine possible combinations of relative work. Equity theory was employed to predict the allocations. For example, if X and Y both tried "very hard" they should have received equal numbers of rewards (10 each). If X tried "very hard" and Y tried "kind of hard," then X deserved more rewards. This equity prediction, however, is ordinal only, not proportional. If X receives more rewards than Y, no matter how many more, ordinal equity is supported. The input adjectives cannot be placed in a ratio.

The only way to differentiate ordinal and proportional equity predictions with such ordinally scaled independent variables is to assign numbers to the adjectives, such as *a little bit* = 1, *kind of hard* = 2, and so on. Assuming such an underlying interval scale, the nine data points in the Anderson and Butzin design that reflect allocation behavior would be spaced differently if the subjects employed an ordinal equity rule than if they employed a proportional equity rule. Granting such a scaling assumption, Anderson and Butzin's data with children are consistent with ordinal equity, whereas Anderson's (1976) earlier data with adults appear to be more consistent with proportional equity.

Finally, in interpreting the studies we have reviewed it is worth bearing in mind that we have stressed children's probable incapacity to generate the logical preconditions for calculating equity ratios. We have thereby assumed that it is more useful to consider equity theory in the proportional sense explicit in Adams's theoretical formulations than in the less strict sense implied by the belief of equity researchers that the theory is corroborated whenever the outcome data show ordinal relationships that are congru-

ent with the theory, even though the data do not attain the exact numerical values that the formal theory suggests. It is difficult to judge whether one should hold equity theory accountable in terms of the standards of the theory or in terms of the (more relaxed) standards of tests of the theory. We have opted for the former over the latter. But it is noteworthy that had we opted for the standard of past empirical tests, then the ordinal equity responses of children between 7 and 13 years of age would be interpreted as supporting equity theory rather than interpreted—as we have done—as not directly supportive.

Future tests of equity in young children would be improved if (a) they were explicit about the degree to which, as a theoretical construct, equity requires cognitive interval scaling and proportional thinking; (b) they incorporated experimental procedures that "unconfound" the capacity to think in ratio terms and the willingness to think this way or to perform behaviors indicative of such thinking; (c) they examined the ways in which the development of allocative skills and the development of logico-mathematical skills are temporally related to each other; and (d) they made sure that behaviors indicative of proportional thinking were not the results of correspondence procedures that do not necessarily require proportional thinking. Equity is a slippery construct that requires explicit formulation and the careful choice of experimental procedures that are carefully tailored to the preferred theoretical formulation. In these respects equity is like other justice constructs. (For a discussion of the construct validity of relative deprivation and the requirements of validation tests, see Cook, Crosby, & Hennigan, 1977.)

Theoretical Implications

Up to now we have made the point that equity theory in the proportional form proposed by Adams (1965) is probably inappropriate when applied to children under the age of 13. We have made this point (a) because on logico-mathematical tasks structurally analogous to equity tasks children do not behave in a manner consistent with the equity assumption that they form cognitive ratios and (b) because the allocative behavior of children in equity experiments is not entirely consistent with a proportional interpretation. However, we have also made the point that future research with better designs and procedures may produce results commensurate with proportional equity theory, though we are skeptical.

The research reported in Tables 3 and 4 suggests a three-step development of proportional thought, as evidenced by children's responses to logico-mathematical problems. Children under approximately 6 years of age do not preserve an ordinal relationship from one dimnesion (e.g., distance traveled) to another (e.g., speed). Middle school children from about 6 to 12 years old preserve such ordinal relationships by setting up corresponding series, but do not relate two dimensions in terms of fractions or ratios. Finally, adolescents and adults above 13 years of age can and do compare ratios for proportionality.

Table 5
Results of Allocation Research in Three Age Groups

Result	Age in years: 3–5	6–12	13+
Self-interest	5	0	0
Equality	11	4	0
Ordinal equity	5	29	4
Proportional equity	0	0	9

Table 5 reports the research results in Table 2 that are consistent with each of the four dominant allocation norms—self-interest, equality, ordinal equity, and proportional equity. The studies are divided into columns according to the ages of the subjects: 3- to 5-year-olds make what appear to be self-interest and equality allocations; 6- to 12-year-olds make what seem to be ordinal equity allocations; and 13-year-olds and older seem to prefer proportional equity allocations. Thus, allocation behaviors seem to follow a sequence of stages similar to those of thought in general and are roughly correspondent to Piaget's sequential stages of

Table 6
Three Steps in the Development of Allocation Behavior

Step	Comparison: Person A	Comparison: Person B	Probable onset age in years
Unidimensional	Outcome A ↔ Outcome B Input A ↔ Input B		3
Ordinal	Outcome A > Outcome B ↕ Input A > Input B		6
Proportional	$\frac{\text{Outcome A}}{\text{Input A}} \leftrightarrow \frac{\text{Outcome B}}{\text{Input B}}$		13

intellectual development: preoperational, concrete operational, and formal operational thought.

Unidimensional allocative comparisons of children under 6 years of age. Although self-interest and equality seem to be very different, they share one attribute not possessed by other allocation principles: Both require comparisons on one dimension only—the reward dimension. In the Leventhal and Anderson (1970) design, the child allocator does not have to recall or consider relative work at all to make a self-interest or equality allocation. Self-interest and equality are simple, unidimensional allocation principles within the logical capacity of the preschool child.

A glance at Table 6 should clarify this unidimensional interpretation. The arrows in this table link the objects or concepts to be compared by the reward allocator. The literature suggests that a 4-year-old, for example, can compare the outcomes of Persons A and B to determine whether they are equal or which is larger. Also, the 4-year-old can compare the inputs of Person A and Person B. Thus, the allocation cognitions of the unidimensional child involve only the observation that outcomes or inputs are the same or different for different persons.

Ordinal equity theory in children under 13 years of age. The middle school child is usually capable of preserving ordinal relationships on two or more dimensions for comparison. In other words, the ordinal equity child is able to compare the comparisons. In terms of Table 6, the child notes first that Person A's outcomes are greater than Person B's outcomes and that Person A's inputs are greater than Person B's inputs. But then the child goes on to compare the relationships to see whether or not they are consistent. If A's inputs are greater than B's, but B's outcomes are greater than A's, then equity is violated. In essence, the child is comparing ranks between people on two dimensions, a comparison of abstractions. (The capacity to hold relationships in mind for comparison is an attribute of Piaget's concrete operational thought. Piaget [1964] discussed this capacity as an element of the ability to seriate, or rank, objects on some dimension such as size or number. To seriate sticks, for example, Piaget argued that it is necessary to simultaneously realize that $A < B$ and $B < C$.)

Ordinal equity, in the allocation context, is nothing new. Indeed, Homans's (1961) distributive justice concept, which he applies to adults, is an ordinal allocation rule:

> As a practical matter, distributive justice is realized when each of the various features of his investments and his activities, put into rank order in comparison with those of other men, falls in the same place in all the different rank orders. (p. 264)

As far as ordinal equity with children is concerned, it may have two substages. The first is ordinal equity, as just described, in which two ordinal comparisons are compared ordinally to see if the ranks differ. The second, perhaps beginning around 9 years of

age, could be called *impending interval equity*. At this point, the child may come to believe that inequity exists, even though Person A has greater outcomes and inputs than Person B. This feeling of inequity would occur when A's inputs were much larger than B's but his outcomes were only a little larger. Piaget, Grize, Szeminska, and Vinh Bang (1968) have explored this onset of measurement in the physical domain; but it has not been explored in allocation research.

Proportional Equity Theory in Adolescents

Formal operational thinkers—normally persons of adolescent age or older—are capable of the proportional thought implied by Adams (1965). Adams's theory implies that the ratios that are compared for proportionality consist of the inputs (or work) and the outcomes (or rewards) of a single person:

$$\frac{\textit{Reward A}}{\textit{Work A}} = \frac{\textit{Reward B}}{\textit{Work B}}.$$

This formulation causes certain cognitive difficulties because the numerator and denominator are expressed in different units. However, the formally equivalent proportion, in which a dimension rather than a person is expressed in the ratio, is cognitively simpler:

$$\frac{\textit{Work A}}{\textit{Work B}} = \frac{\textit{Reward A}}{\textit{Reward B}}.$$

It is probably comprehended earlier by the adolescent. When more than two persons are being compared, the simplier formulation can be modified to take the form

$$\frac{\textit{Work A}}{\textit{Work A} + \textit{Work B} + \textit{Work C}} = \frac{\textit{Reward A}}{\textit{Reward A} + \textit{Reward B} + \textit{Reward C}}.$$

We suggest that subjects cognitively assess each person in terms of whether his work, as a proportion of all the work, is equivalent to his reward, as a proportion of all the reward. Hook (1978) found this to be the way 13-year-old proportional thinkers made allocations; and other than for Homans, it is the formulation adopted by the allocation theorists who preceded Adams (see Patchen, 1961; Sayles, 1958). It is also consistent with work by integration theorists (Anderson & Farkas, 1975) on the relative accuracy of prediction of the three equity equations just mentioned.

Reference Notes

1. Vinacke, W. E., & Stanley, S. *Strategy in a masculine quiz game* (Tech. Rep. 2). Honolulu: University of Hawaii, 1962.
2. Staub, E. *The effects of success and failure on children's sharing behavior.* Paper presented at the meeting of the Eastern Psychological Association, Washington, D.C., April 1963.
3. Antone, D., & Hendricks, M. *Children, equity, and distortion of relative inputs.* Unpublished manuscript, Northwestern University, 1976.
4. Hook, J., Brockett, N., & Smith, D. *The development of equity and altruism in judgments of reward and damage allocation.* Unpublished manuscript, University of Nebraska—Lincoln, 1979.
5. Cohen, E., & Sampson, E. *Distributive justice: A preliminary study of children's equal and equitable allocations of reward using the doll play technique.* Paper presented at the meeting of the Eastern Psychological Association, New York, April 1975.
6. Steffe, L., & Parr, R. *The development of the concepts of ratio and fraction in the fourth, fifth, and sixth years of the elementary school* (Tech. Rep. 49). Madison: University of Wisconsin, Research and Development Center for Cognitive Learning, 1968.

References

Adams, J. S. Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology.* New York: Academic Press, 1965.

Anderson, N. Equity judgments as information integration. *Journal of Personality and Social Psychology,* 1976, *33,* 291–299.

Anderson, N., & Butzin, C. Integration theory applied to children's judgments of equity. *Developmental Psychology,* 1978, *14,* 593–606.

Anderson, N. H., & Farkas, A. Integration theory applied to models of inequity. *Personality and Social Psychology Bulletin,* 1975, *1,* 588–591.

Benton, A. Productivity, distributive justice, and bargaining among children. *Journal of Personality and Social Psychology,* 1971, *18,* 68–78.

Berkowitz, L. Social norms, feelings, and other factors affecting helping behavior and altruism. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 6). New York: Academic Press, 1972.

Brainerd, C. *Piaget's theory of intelligence*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Brainerd, C., & Allen, T. Training and generalization of density conservation: Effects of feedback and consecutive similar stimuli. *Child Development,* 1971, *42,* 693–704.

Brickman, P., & Bryan, J. H. Moral judgment of theft, charity, and third-party transfers that increase or decrease equality. *Journal of Personality and Social Psychology,* 1975, *31,* 156–161.

Brickman, P., & Bryan, J. H. Equity versus equality as factors in children's moral judgments of thefts, charity, and third-party transfers. *Journal of Personality and Social Psychology,* 1976, *34,* 757–761.

Bruner, J. S., & Kenney, H. On relational concepts. In J. S. Bruner, R. R. Olver, & P. M. Greenfield, *Studies in cognitive growth*. New York: Wiley, 1966.

Bryan, J. H. Why children help: A review. *Journal of Social Issues,* 1972, *28,* 87–101.

Campbell, D. T. On the genetics of altruism and counterhedonic components in human culture. *Journal of Social Issues,* 1972, *28,* 21–38.

Chapman, R. The development of children's understanding of proportions. *Child Development,* 1975, *46,* 141–148.

Cohen, R. J. Mastery and justice in laboratory dyads: A revision and extension of equity theory. *Journal of Personality and Social Psychology,* 1974, *29,* 464–474.

Cook, T. D., Crosby, F., & Hennigan, K. M. The construct validity of egoistic relative deprivation. In R. L. Miller & J. R. Surls (Eds.), *Comparison processes: Theoretical and empirical perspectives.* Washington, D.C.: Hemisphere, 1977.

Coon, R., Lane, I., & Lichtman, R. J. Sufficiency of reward and allocation behavior. *Human Development,* 1974, *17,* 301–313.

Crockenberg, S., Bryant, B., & Wilce, L. S. The effects of cooperatively and competitively structured learning environments of inter- and intrapersonal behavior. *Child Development,* 1976, *47,* 386–396.

Damon, W. Early conceptions of positive justice as related to the development of logical operations. *Child Development,* 1975, *46,* 301–312.

Dreman, S. B. Sharing behavior in Israeli school children: Cognitive and social learning factors. *Child Development,* 1976, *47,* 186–194.

Fischbein, E., Pampu, I., & Manzat, I. Comparison of ratios and the chance concept in children. *Child Development,* 1970, *41,* 377–389.

Garrett, J., & Libby, W. Role of intentionality in mediating responses to inequity in the dyad. *Journal of Personality and Social Psychology,* 1973, *28,* 21–27.

Handlon, B., & Gross, P. The development of sharing behavior. *Journal of Abnormal and Social Psychology,* 1959, *59,* 425–428.

Homans, G. C. *Social behavior: Its elementary forms.* New York: Harcourt, Brace & World, 1961.

Hook, J. G. The development of equity and logico-mathematical thinking. *Child Development,* 1978, *49,* 1035–1044.

Inhelder, B., & Piaget, J. *The growth of logical thinking*. London: Routledge & Kegan Paul, 1958.

Kahn, A. Reactions to generosity or stinginess from an intelligent or stupid partner: A test of equity theory in a direct exchange relationship. *Journal of Personality and Social Psychology,* 1972, *21,* 116–123.

Kidder, L., Bellettirie, G., & Cohn, E. Secret ambitions and public performances: The effects of anonymity on reward allocations made by men and women. *Journal of Experimental Social Psychology,* 1977, *13,* 70–80.

Lane, I., & Coon, R. Reward allocation in preschool children. *Child Development,* 1972, *43,* 1382–1389.

Lane, I., Messe, L., & Phillips, J. Differential inputs as a determinant in selection of a distributor of rewards. *Psychonomic Science,* 1971, *22,* 228–229.

Larsen, G., & Kellogg, J. A developmental study of relations between conservation and sharing behavior. *Child Development,* 1974, *45,* 849–851.

Lawler, E. E. Effects of hourly overpayment on productivity and work quality. *Journal of Personality and Social Psychology,* 1968, *10,* 306–313.

Lee, L. C. The concomitant development of cognitive and moral modes of thought: A test of selected deductions from Piaget's theory. *Genetic Psychology Monographs,* 1971, *83,* 93–146.

Lerner, M. The justice motive: "Equity" and "parity" among children. *Journal of Personality and Social Psychology,* 1974, *29,* 539–550.

Leventhal, G., & Anderson, D. Self-interest and maintenance of equity. *Journal of Personality and Social Psychology,* 1970, *15,* 57–62.

Leventhal, G., & Lane, D. Sex, age, and equity behavior. *Journal of Personality and Social Psychology,* 1970, *15,* 312–316.

Leventhal, G., & Michaels, J. Extending the equity model: Perception of inputs and allocation of reward as a function of duration and quantity of performance. *Journal of Personality and Social Psychology,* 1969, *12,* 303–309.

Leventhal, G., Michaels, J., & Sanford, C. Inequity and interpersonal conflict. *Journal of Personality and Social Psychology,* 1972, *23,* 88–102.

Leventhal, G., Popp, A., & Sawyer, L. Equity or equality in children's allocation of reward to other persons? *Child Development,* 1973, *44,* 753–763.

Leventhal, G., Weiss, T., & Buttrick, R. Attribution of value, equity, and the prevention of waste in reward allocation. *Journal of Personality and Social Psychology,* 1973, *27,* 276–286.

Leventhal, G., Weiss, T., & Long, G. Equity, reciprocity, and reallocating rewards in the dyad. *Journal of Personality and Social Psychology,* 1969, *13,* 300–305.

Libby, W., & Garrett, J. Role of intentionality in mediating children's responses to inequity. *Developmental Psychology*, 1974, *10*, 294–297.
Long, G., & Lerner, M. Deserving, the "personal contact," and altruistic behavior by children. *Journal of Personality and Social Psychology*, 1974, *29*, 551–556.
Lovell, K. A follow-up study of Inhelder and Piaget's *The growth of logical thinking*. *British Journal of Psychology*, 1961, *2*, 143–153.
Lovell, K., & Butterworth, I. Abilities underlying the understanding of proportionality. *Mathematics Teaching*, 1966, *37*, 5–9.
Lunzer, E. A. Problems of formal reasoning in test situations. *Monographs of the Society for Research in Child Development*, 1965, *30*(2, Serial No. 100).
Masters, J. C. Effects of social comparison upon subsequent self-reinforcement behavior in children. *Journal of Personality and Social Psychology*, 1968, *10*, 391–401.
Miller, D. T., & Smith, J. The effect of own deservingness and deservingness of others on children's helping behavior. *Child Development*, 1977, *48*, 617–620.
Morgan, W. K., & Sawyer, J. Bargaining, expectation, and the preference for equity. *Journal of Personality and Social Psychology*, 1967, *6*, 139–149.
Nelson, S., & Dweck, C. Motivation and competence as determinants of young children's reward allocation. *Developmental Psychology*, 1977, *13*, 192–197.
Olejnik, A. B. The effects of reward-deservedness on children's sharing. *Child Development*, 1976, *47*, 380–385.
Patchen, M. *The choice of wage comparisons*. Englewood Cliffs, N.J.: Prentice-Hall, 1961.
Peterson, C., Peterson, J., & McDonald, N. Factors affecting reward allocations by preschool children. *Child Development*, 1975, *46*, 942–947.
Piaget, J. Logique et équilibre dans les comportements du sujet. In L. Apostel, et al. (Eds.), *Logique et équilibre*. Paris: Presses Universitaires de France, 1957.
Piaget, J. *The child's conception of number*. London: Routledge & Kegan Paul, 1964.
Piaget, J. *The child's conception of movement and speed*. New York: Basic Books, 1970.
Piaget, J. *Understanding causality*. New York: Norton, 1974.
Piaget, J., Grize, J., Szeminska, A., & Vinh Bang. *Epistemologie et psychologie de la function*. Paris: Presses Universitaires de France, 1968.
Piaget, J., & Inhelder, B. *The child's conception of space*. London: Routledge & Kegan Paul, 1956.
Piaget, J., & Inhelder, B. *The origin of the idea of chance in children*. New York: Norton, 1975.
Piaget, J., Inhelder, B., & Szeminska, A. *The child's conception of geometry*. New York: Basic Books, 1960.
Reis, H. T., & Gruzen, J. On mediating equity, equality, and self-interest: The role of self-presentation in social exchange. *Journal of Experimental Social Psychology*, 1976, *12*, 487–503.
Sampson, E. E. Studies of status congruence. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 4). New York: Academic Press, 1969.
Sayles, L. R. *Behavior of industrial work groups: Prediction and control*. New York: Wiley, 1958.
Schopler, J. An attributional analysis of some determinants of reciprocating a benefit. In J. Macauley & L. Berkowitz (Eds.), *Altruism and helping behavior*. New York: Academic Press, 1970.
Shapiro, E. Effect of expectations of future interactions on reward allocations in dyads: Equity or equality? *Journal of Personality and Social Psychology*, 1975, *31*, 873–880.
Smith, R. J., & Cook, P. E. Leadership in dyadic groups as a function of dominance and incentives. *Sociometry*, 1973, *36*, 561–568.
Staub, E. Instigation to goodness: The role of social norms and interpersonal influence. *Journal of Social Issues*, 1972, *28*, 131–150.
Streater, A., & Chertkoff, J. Distribution of rewards in a triad: A developmental test of equity theory. *Child Development*, 1976, *47*, 800–805.
Tomlinson-Keasey, C., & Keasey, C. The mediating role of cognitive development in moral judgment. *Child Development*, 1974, *45*, 291–298.
Tompkins, B., & Olejnik, A. Children's reward allocations: The impact of situational and cognitive factors. *Child Development*, 1978, *49*, 526–529.
Uesugi, T. K., & Vinacke, W. E. Strategy in a feminine game. *Sociometry*, 1963, *26*, 75–88.
Ugurel-Semin, R. Moral behavior and moral judgment of children. *Journal of Abnormal and Social Psychology*, 1952, *47*, 463–474.
Von Grumbkow, J., Deen, E., Steensma, H., & Wilke, H. The effect of future interaction on the distribution of rewards. *European Journal of Social Psychology*, 1976, *6*, 119–123.
Webb, R. A. Concrete and formal operations in very bright six to eleven year olds. *Human Development*, 1974, *17*, 292–300.
Wright, B. Altruism in children and the perceived conduct of others. *Journal of Abnormal and Social Psychology*, 1942, *37*, 218–233. (a)
Wright, B. The development of the ideology of altruism and fairness in children. *Psychological Bulletin*, 1942, *39*, 485–486. (b)

Received January 31, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 3, 446–461

Group Reaction Time Distributions and an Analysis of Distribution Statistics

Roger Ratcliff
Dartmouth College

A method of obtaining an average reaction time distribution for a group of subjects is described. The method is particularly useful for cases in which data from many subjects are available but there are only 10–20 reaction time observations per subject cell. Essentially, reaction times for each subject are organized in ascending order, and quantiles are calculated. The quantiles are then averaged over subjects to give group quantiles (cf. Vincent learning curves). From the group quantiles, a group reaction time distribution can be constructed. It is shown that this method of averaging is exact for certain distributions (i.e., the resulting distribution belongs to the same family as the individual distributions). Furthermore, Monte Carlo studies and application of the method to the combined data from three large experiments provide evidence that properties derived from the group reaction time distribution are much the same as average properties derived from the data of individual subjects. This article also examines how to quantitatively describe the shape of reaction time distributions. The use of moments and cumulants as sources of information about distribution shape is evaluated and rejected because of extreme dependence on long, outlier reaction times. As an alternative, the use of explicit distribution functions as approximations to reaction time distributions is considered.

Despite the recent popularity of reaction time research, the use of reaction time distributions for both model testing and model development has been largely ignored. This is surprising in view of the fact that properties of distributions can prove decisive in discriminating among models (Sternberg, Note 1) and can falsify models that quite adequately describe the behavior of mean reaction time (Ratcliff & Murdock, 1976).

Two methods have been used to obtain distributional or shape information. One method, advocated by Sternberg (1969; Sternberg, Note 2), is to use moments and cumulants to describe distribution shape without assuming any particular reaction time distribution function. A second method, used by Ratcliff and Murdock, is to assume an explicit distribution function and use the parameters of this distribution to provide information about shape. Both these methods are unattractive because they require 5 to 10 times the number of observations usually collected in an experiment. For example, to fit an explicit function such as a gamma distribution to experimentally obtained reaction times, a minimum of about 100 observations per subject per condition are required for reliable convergence of fitting procedures and stability of parameter estimates. Similarly, to obtain stable estimates of higher moments, several thousand observations per condition are typically required. The necessity for a large number of observations becomes a particular problem in experimental endeavors in

This research was supported by Grant APA 146 from the National Research Council of Canada and Grant OMHF 164 from the Ontario Mental Health Foundation to B. B. Murdock, Jr.

I would like to thank Ben Murdock for his support, help, and criticism. I also wish to thank David Andrews for many useful suggestions and Gail McKoon for help in making the article somewhat understandable.

Requests for reprints should be sent to Roger Ratcliff, Department of Psychology, Dartmouth College, Hanover, New Hampshire 03755.

which the test materials used require a great deal of time and effort for construction (e.g., paragraphs; Kintsch, 1974). For such research programs, it would take years to construct enough materials to allow application of either of the two distributional methods.

In the first part of this article, I present a method for combining data from individual subjects to produce group reaction time distributions based on as few as 10 observations per subject cell. To form group distributions, reaction times for each subject are organized in ascending order, and quantiles are calculated. The quantiles are then averaged over subjects to give group quantiles (Vincent averaging; Vincent, 1912). From the group quantiles, a group reaction time distribution can be constructed. This group distribution method averages over individual subjects' data in a way that retains shape information, and this is demonstrated in three ways: First, it is shown that for certain distributional forms (exponential, Weibull, and logistic), Vincent averaging of individual distributions of a particular form with different parameters results in a group distribution of the same functional form. Second, a distribution that has been used to describe reaction time data (Ratcliff & Murdock, 1976) was used in Monte Carlo studies to generate reaction times that were then combined according to Vincent's method. With 20 reaction times per pseudosubject, the group distributions generated by this method have the same form as the distribution used to generate the data. Third, the method was applied to the combined data from three large recognition memory experiments that used the study–test procedure (Ratcliff & Murdock, 1976), with about 120,000 observations in total. It is shown that parameters derived from fitting a distribution function (used by Ratcliff & Murdock, 1976) to the group distribution are the same as averages of the parameters derived from fitting the function to individual distributions.

In the second part of the article, I critically examine the use of moments and cumulants for describing distribution shape. The stability of moment and cumulant estimates is examined first by calculating sampling standard deviations and second by observing the stability of estimates when outlier reaction times are trimmed from the distribution. In addition, the use of empirical distribution functions to provide information about distribution shape is examined.

The notion of shape can be defined in different ways. Mosteller and Tukey (1977, chap. 1) defined shape as what is left when location (position of the distribution on the abscissa) and scale (the scale on the abscissa) are given up. They showed that shape cannot be defined in terms of the mathematical form of the distribution function. For example, the family of beta density functions have the same functional form, but differ widely in shape (Mosteller & Tukey, 1977, p. 9). However, one of the most striking properties of reaction time distributions is that in the main they all have roughly the same shape, being skewed to the right. (Occasionally normality is claimed for simple reaction time distributions, but this is probably not true [Mosteller, & Tukey, 1977, p. 11].) The group distribution method is concerned with averaging over subjects while preserving distribution shape, which for distribution functions shaped like reaction time distributions often turns out to be much the same as preserving the functional form, as is shown later.

Reaction time distributions have been examined in some detail with respect to specific mathematical models. McGill (1963) provided an excellent summary of work prior to 1963 and presented formal theory for a number of latency models. Green and Luce (1971) have used transform techniques in conjunction with a specific decision model to decompose reaction time distributions into component distributions, and this method of decomposition has been used in testing a neural timing theory (Luce & Green, 1972). Hohle (1965), Snodgrass (1969), and Snodgrass, Luce, and Galanter (1967) have fitted various empirical distributions to choice and simple reaction time data. None of this work, however, provides a general approach to obtaining distributional or shape information.

Before proceeding to a discussion of methods, I briefly illustrate potential uses of distributional information by listing predictions made by four models about distribution shape. First, serial scanning models of item recognition that assume independent and identically distributed comparison stages predict (by the central limit theorem) that as the number of comparison stages increases, the skewness of the reaction time distribution will decrease, and so the distribution will become more normal in shape. Second, the Atkinson and Juola (1973) model of item recognition predicts bimodal reaction time distributions. Third, the multiple observations model for signal detection (Pike, 1973) predicts that when the count criteria increase, mean latency will increase and skewness will decrease. Fourth, the random walk model for item recognition (Ratcliff, 1978) predicts that as the relatedness between probe item and memory item decreases, the mode and mean of the reaction time distribution will diverge. These examples are meant only to indicate the kinds of predictions that models produce and thus the kinds of tests for which distributional analyses prove useful.

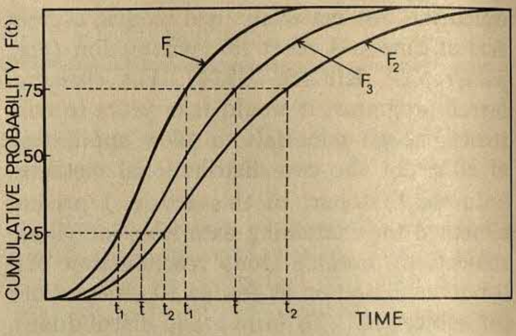

Figure 1. An example of Vincent (1912) averaging applied to cumulative distribution functions; $F_3(t)$ is the Vincent average curve of the curves $F_1(t)$ and $F_2(t)$, and $\bar{t}$ is the average of t_1 and t_2.

Group Reaction Time Distributions

In experimental psychology it is usual to generalize findings across subjects. Often this is done by averaging data over subjects and making inferences based on the group data. Unfortunately, if raw reaction times from several subjects were simply combined to obtain distributional information, then the group distribution would not reflect the shape of the individual distributions. As an illustration, consider two subjects' unimodal reaction time distributions with respective means of 500 msec and 900 msec, each with 100-msec standard deviations. Simply combining the data would give a bimodal distribution, and this would not reflect the unimodal, individual distributions.

If there are enough observations per subject cell, then the best way to obtain distribution information is to derive distributional or shape estimates for each subject cell and then average these estimates over subjects. For example, Ratcliff and Murdock (1976) have used the function arising from the convolution of the normal $[N(\mu,\sigma)]$ and exponential $[g(t) = (1/\tau)e^{-t/\tau}]$ distribution functions, $f(t)$, as an empirical summary of the shape of individual subjects' reaction time distributions (see also Hohle, 1965). Generalization was then accomplished by finding the average of each convolution parameter (μ, σ, and τ) across subjects. The expression for the convolution is

$$f(t) = \frac{e^{-[(t-\mu)/\tau]+\sigma^2/2\tau^2}}{\tau(2\pi)^{\frac{1}{2}}} \times \int_{-\infty}^{[(t-\mu)/\sigma]-\sigma/\tau} e^{-y^2/2}dy. \quad (1)$$

In a similar vein, Sternberg (Note 2) found four cumulants of distributions for each subject cell and generalized across subjects by averaging each cumulant across subjects for each cell. However, the usual experiment does not provide the number of observations required for these methods. Distribution information can still be obtained by using the group distribution method.

Group Distribution Method

The method is very similar to the technique devised by Vincent (1912) for plotting learning curves. In Vincent's procedure, each individual's learning curve is divided into equal fractions (number of trials to 10%, 20%, . . .), and performance of subjects at each fraction is summed and then

averaged. Figure 1 shows an example of this "Vincentizing" procedure applied to two cumulative reaction time distributions to produce the average cumulative distribution. In essence, reaction times at a fixed probability level (quantile) from the two distributions are averaged to give the mean quantile reaction time.

The procedure for estimating the sample quantiles is carried out as follows: Each subject's reaction times $T_1, \ldots, T_n$ are arranged in ascending order of magnitude: $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$, where $T_{(i)}$ is the ith order statistic (David, 1970; Sarhan & Greenberg, 1962). From these ordered reaction times, q sample quantiles are estimated for each subject's data (generally with $q < n$). Each quantile is then averaged across subjects to give a mean $m\%$ sample quantile. In detail, suppose there are n observations for the first subject and one wishes to obtain q quantiles ($q < n$). Then for each subject, each ordered latency $T_{(i)}$ is replaced by q equal latencies, $T_{(i)}$, thereby forming a list that is the length of the product of q and n: $T_{(1)}, T_{(1)}, \ldots, T_{(1)}, T_{(2)}, T_{(2)}, \ldots, T_{(2)}, T_{(3)}, \ldots$. To calculate the first quantile, the first n latencies are summed and divided by n; the second quantile is given by the sum of the next n latencies divided by n, and so on. This procedure is equivalent to simple linear interpolation. For example, if there were 14 responses and deciles were to be calculated, the first decile would be given by $(10/14)T_{(1)} + (4/14)T_{(2)}$, the second by $(6/14)T_{(2)} + (8/14)T_{(3)}$, the third by $(2/14)T_{(3)} + (10/14)T_{(4)} + (2/14)T_{(5)}$, and so on. When the quantiles have been calculated for each subject, each quantile is averaged across subjects to give group quantiles.

Group distribution histograms can be constructed by plotting quantiles on the abscissa and then constructing rectangles between adjacent quantiles such that all the rectangles have equal areas, as in Figure 2.

Several points about this method need discussion. First, the group distribution should be thought of as representing the distribution of the average subject, just as average reaction time represents the reaction time of the average subject. Second, order statistics are biased estimators of quantiles

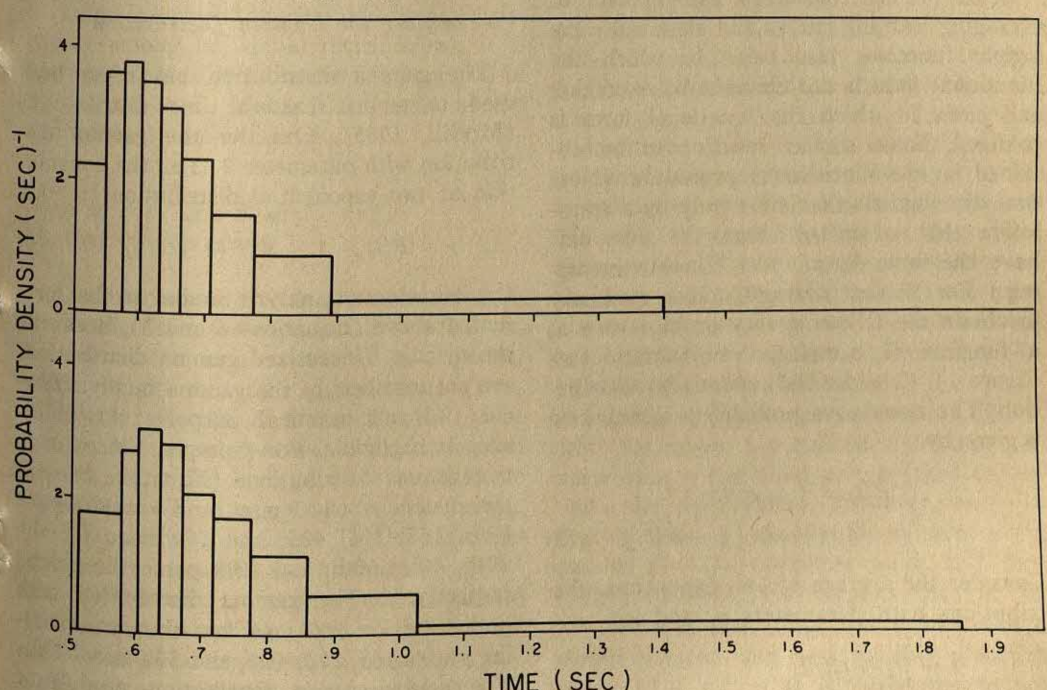

Figure 2. Two sample group reaction time distributions for 10% quantiles. (Data are from the three experiments reported later and represent correct rejections in Output Blocks 1 and 4.)

(David, 1970, chap. 4). However, if there are roughly the same number of observations for each subject cell and if the individual distributions for each subject cell have approximately the same shape, then the group distribution will reflect the same bias as the individual distributions. Third, it is a general problem that the shape of a group curve may not reflect the shape of individual curves. This problem was considered in great detail in the mid-1950s (Bakan, 1954; Estes, 1956; Hayes, 1953; Sidman, 1952; Spence, 1956). The general conclusion reached was that group curves often do not reflect the form of individual curves but that if care is taken group curves can be used to test hypotheses about individual curves. The next two sections examine this problem with respect to the Vincent averaging procedure, and later sections examine the problem with respect to Vincent averaging of reaction time distributions.

Some Exact Results for Vincentized Curves

Estes (1956) considered the problem of averaging learning curves and classified some simple functions into cases in which the functional form is not changed by averaging and cases in which the functional form is changed. Some similar results can be obtained for the Vincentizing procedure. (Note that distributions that differ only by a translation that is shifted along the time axis have the same form under Vincent averaging.) For Vincent averaging, it is necessary to obtain the following relationship: time as a function of cumulative probability (see Figure 1). Consider the exponential distribution. The cumulative probability distribution is given by

$$F(t) = 1 - e^{-t/\tau};$$

$$\therefore \qquad t = -\tau \ln [1 - F(t)]. \qquad (2)$$

Consider the average of two exponential distributions with parameters τ_1 and τ_2:

$$\bar{t} = (t_1 + t_2)/2$$

$$= -\frac{(\tau_1 + \tau_2)}{2} \ln [1 - F(t)]. \qquad (3)$$

Thus the "Vincentized" average of n exponential distributions is exponential with parameter $\Sigma^n{}_{i=1}\ \tau_i/n$. For the Weibull distribution,

$$F(t) = 1 - e^{-(t/\tau)^\gamma},$$

with fixed parameter γ; the Vincentized distribution is also a Weibull distribution, with parameter $\Sigma^n{}_{i=1}\ \tau_i/n$. Similarly, for the logistic distribution,

$$F(t) = 1/[1 - e^{-(t-\alpha)/\beta}];$$

the Vincentized distribution is also logistic, with parameters $\alpha = \Sigma^n{}_{i=1}\ \alpha_i/n$ and $\beta = \Sigma^n{}_{i=1}\ \beta_i/n$. Although normal distributions will not give exactly a normal distribution when Vincentized, the logistic distribution is a very good approximation to the normal distribution, so that any differences are probably very small. It should be noted that the exponential, Weibull ($\gamma > 1$), and logistic distributions have been postulated (or are very similar to distributions that have been postulated) to represent the distributions of processing stages in various models.

Vincentizing the Gamma Distribution

The gamma distribution has often been used to model reaction time distributions (McGill, 1963). Consider the gamma distribution with parameter 2 (i.e., the convolution of two exponential distributions):

$$F(t) = 1 - e^{-t/\tau}(1 + t/\tau). \qquad (4)$$

By following an analysis similar to that presented above (Equations 2 and 3), it can be shown that Vincentized gamma distributions are not members of the gamma family. However, for all practical purposes the difference is negligible. For example, Vincentizing two gamma distributions (Equation 4) with parameters $\tau = 100$ msec and $\tau = 300$ msec gives 165, 276, 405, and 599 msec for the 20th, 40th, 60th, and 80th percentile points, respectively. The gamma distribution with parameter $\tau = 200$ msec has the corresponding points 165, 275, 404, and 599 msec. Thus Vincentizing gamma distributions produces a distribution that is very similar in shape to another gamma distribution.

The examples so far have all considered the Vincentizing of combinations of distributions that differ from one another only in the parameters that have dimensions of time, that is, parameters that represent the duration of some processing stage. There are other parameters that do not represent durations, for example, the number of convolved exponential distributions in the gamma distribution and γ in the Weibull distribution. Vincentizing distributions that vary in these parameters may not produce a distribution that is anything like the average subject's distribution. An extreme example of a distribution with this problem is the beta distribution. Mosteller and Tukey (1977, p. 9), in considering the problems involved in dealing with distribution shape, presented a figure showing the family of beta distributions to illustrate that even distributions belonging to the same family can differ widely in shape. From the graphs presented in Mosteller and Tukey, it can be seen that very serious problems may be involved in averaging across distributions of widely differing shape. To decide whether Vincent averaging will work in cases in which distribution shape varies widely among individual distributions, it is probably best to test the method as above or to perform some Monte Carlo tests as described in the next section.

Some Monte Carlo Studies Using the Convolution Model

The distribution that is the convolution of the normal and the exponential distributions (Equation 1) has been used as an empirical model of reaction time distributions (Ratcliff, 1978; Ratcliff & Murdock, 1976). The fits of the convolution to the data are good enough to make it reasonable to use the convolution in Monte Carlo studies testing the Vincentizing procedure. The Monte Carlo studies are presented to illustrate the use of the Vincentizing procedure under optimal conditions in which the form of the individual distributions is known.

To use the Monte Carlo method it is necessary to generate a random number from the convolution of normal and exponential distributions. This can be accomplished by simply adding a random number generated from the normal distribution and a random number generated from the exponential distribution. Most computer systems have a random number generator that will produce random numbers between 0 and 1 from a rectangular distribution. Equation 2 can be used to produce exponentially distributed random numbers (with parameter τ) by substituting rectangularly distributed random numbers (RND) for $F(t)$. Normally distributed random numbers with mean μ and standard deviation σ can be obtained using the method proposed by Box and Muller (1958), as shown in Equation 5:

$$t = [-2\,ln(RND)]^{\frac{1}{2}} \cos(2\pi RND)\sigma + \mu. \quad (5)$$

Each Monte Carlo study consisted of several experiments (typically 50 to 100). In each experiment, 20 reaction times were obtained from each of 40 pseudosubjects. The 20 reaction times were arranged in ascending order and then averaged across subjects to give group 5% quantiles. The convolution model was then fitted to the set of 5% quantiles (5%, 10%, 15%, . . .) using the maximum likelihood method described in Ratcliff and Murdock (1976). Note that the quantile reaction times are derived from random variables; and so, strictly speaking, the parameter estimates do not have the nice properties of maximum likelihood estimators. However, estimating parameters this way is no worse than estimating parameters by, say, the least squares method, because the quantile reaction times are not independent and the expression being fitted is nonlinear. Results are shown in Table 1.

In general, the parameters μ and τ derived from fits to the Vincentized distribution are very close to the input values (used to generate the pseudodata). However, as τ increases (from .05 to .30), the value of σ (input value = .04) becomes more and more underestimated. This suggests that in any practical use, the value of σ is likely to be underestimated and less reliable than the values of μ and τ. It is interesting to note that the values of s_μ, s_σ, and s_τ are very close to the asymptotic variance estimates for the

Table 1
Monte Carlo Studies for the Convolution Model

Input parameter			Fitted parameter and standard error estimate						No. of experiments
μ	σ	τ	μ	s_μ	σ	s_σ	τ	s_τ	
.50	.04	.05	.5005	.0004	.0371	.0003	.0486	.0005	65
.50	.04	.15	.4996	.0004	.0325	.0004	.1498	.0006	101
.50	.04	.30	.5016	.0008	.0275	.0007	.2955	.0015	58/109[a]
.50	.10	.50	.5002	.0014	.0749	.0012	.4994	.0024	98
.50	.04	.055–.250[b]	.4980	.0006	.0314	.0006	.1545	.0012	52

Note. s = the standard error in the mean $([\Sigma(X_i - M)^2/n(n-1)]^{\frac{1}{2}})$.
[a] 51 of the 109 experiments terminated with the fitted value of σ equal to zero.
[b] The 40 pseudosubjects had different τs, ranging from .055 to .250 in steps of .005, M = .1525.

convolution model presented in Ratcliff and Murdock (1976, Table 2). The last series of Monte Carlo experiments presented in Table 1 used 40 pseudosubjects with different τ values. The value of the average Vincentized τ was almost equal to the average input τ. This result shows that the Vincent-averaging properties of the exponential distribution carry over to parameter τ of the convolution to a good approximation.

These Monte Carlo studies show that application of Vincent's (1912) procedure to the distribution that is the convolution of normal and exponential distributions, a distribution that fits response latency distributions reasonably well, introduces little bias into parameter estimates.

Practical Test of the Group Distribution Method

To provide a stable data base for a practical examination of the method, three experiments (with four subjects per experiment) were combined, giving about 120,000 reaction times in total. The experimental procedure was the study–test recognition memory paradigm. The experiments have been reported as Experiments 2 and 1 in Ratcliff and Murdock (1976) and Experiment 1 in Ratcliff (1978); they are referred to here as Experiments 1, 2, and 3, respectively. A brief description of the study–test procedure is presented here; for further details, consult Ratcliff and Murdock.

In each of the three experiments, a list of study words was presented to the subject at about one word per sec, followed by a test list containing all the study words plus an equal number of new words in random order. For each word in the test list, the subject had to respond on a 6-point confidence scale ranging from *sure old* to *sure new*. Study lists were 16 words long, and test lists were 32 words long, except in Experiment 1 in which the study list contained 15 words and the test list 30 words. The list words were randomly sampled from the Toronto word pool (Okada, 1971). Repetitions of words were prohibited until at least two lists had intervened. The test list was self-paced, and words stayed in view until a response was made. In Experiment 1 and Experiment 3 rate of presentation of the study lists was varied between .5 sec and 2 sec per item. Effects on mean reaction time were small, on the order of 40 msec. In the following analyses, data from the different presentation-rate conditions are combined; this does not significantly affect distribution shape.

The experimental data are classified into eight cells, four output- or test-position blocks (2–8, 9–16, 17–24, and 25–32) for high-confidence hits, and the same four output-position blocks for correct rejections. (The first output position is excluded because this reaction time is typically several hundred msec slower than other reaction times in the test list.) This division of data gives about 1,200 observations for each of the 96 subject cells (12 subjects × 8 cells).

It was noted earlier that to test the group distribution method, properties derived from

Table 2
Convolution Model Fits to the Group Reaction Time Distributions and the Average Parameter Values From Fits to Individual Subject Distributions

Output block	Parameter value averaged over subjects			Group distribution value		
	μ	σ	τ	μ	σ	τ
			Hit: $T < 5$ sec			
1	492	38	178	488	36	179
2	498	36	200	494	33	196
3	506	37	231	503	35	225
4	517	41	261	507	37	256
			Correct rejection: $T < 5$ sec			
1	517	37	213	513	36	216
2	523	40	236	519	38	243
3	526	41	272	521	38	273
4	524	43	300	518	41	302
			Hit: $T < 2$ sec			
1	497	41	158	495	39	157
2	502	38	175	499	36	172
3	513	40	192	510	39	186
4	525	45	210	517	42	205
			Correct rejection: $T < 2$ sec			
1	523	40	185	520	38	186
2	530	43	200	525	41	202
3	536	45	218	533	44	223
4	538	49	225	532	48	229

Note. Data are truncated at 5 sec and at 2 sec.

the group distribution must be compared with the averages of the properties derived from the individual distributions. The properties chosen for comparison were the parameters of the convolution model, μ, σ, and τ (see Equation 1). Estimates of these parameters were obtained from the group distributions (2% quantiles) for each of the eight cells by fitting the convolution model to the group quantiles. (See Ratcliff & Murdock, 1976, for the maximum likelihood method of fitting.) Estimates of the parameters were obtained from the individual subject distributions by first fitting the convolution to each subject's distribution and then averaging the obtained estimates over subjects. The estimates given by the two procedures can be compared in Table 2: For two conditions each, estimates with latencies longer than 5 sec eliminated and estimates with latencies longer than 2 sec eliminated. It can be seen that the two procedures give almost identical estimates of the convolution parameters. This supports the claim made earlier that the group distribution provides an unbiased summary of individual data.

Figures 2, 3, and 4 show some sample data. Figure 2 shows group reaction time distributions for correct rejections in Output Block 1 and in Output Block 4. Figure 4 shows group reaction time distributions and fits of the convolution model for hits in all four output blocks. The Figure 2 distributions are based on 20% quantiles, the Figure 4 distributions on 2% quantiles. Figure 3 shows some sample fits of the convolution model to reaction time distributions for individual subjects for hits in Output Block 1. Although the chi-squares are often significant (because the large numbers of observations make the

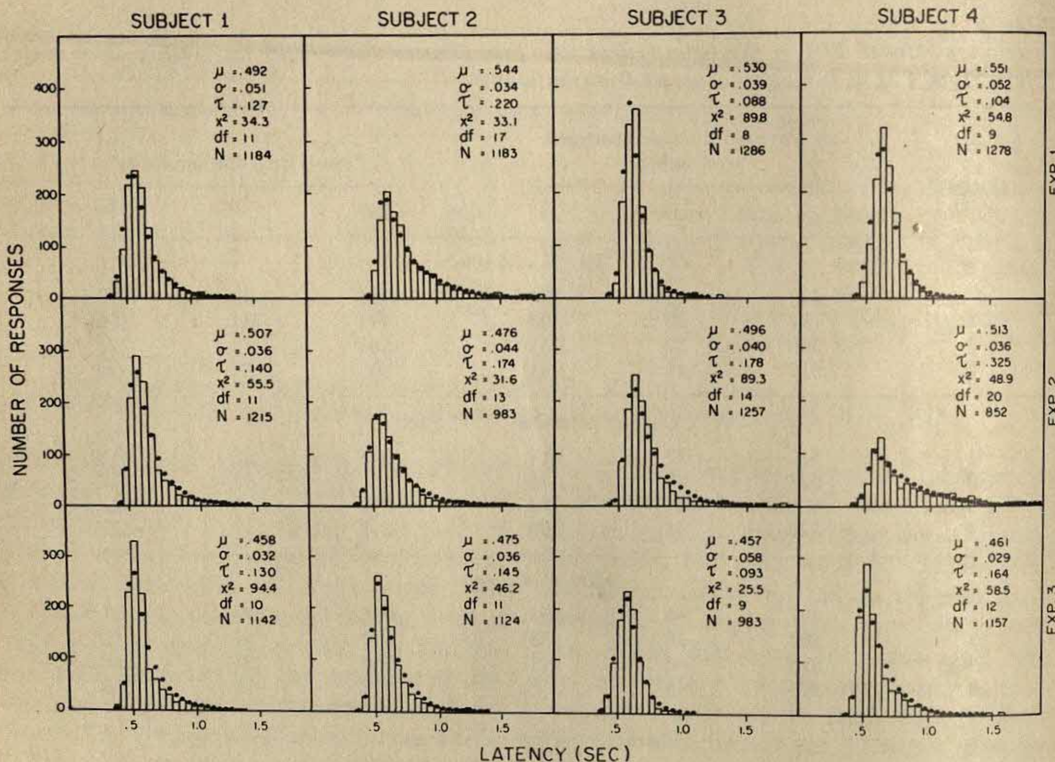

Figure 3. Empirical and fitted latency distributions for hits, Output Block 1. (Exp. 1 = Experiment 2, Ratcliff and Murdock, 1976; Exp. 2 = Experiment 1, Ratcliff and Murdock, 1976; Exp. 3 = Experiment 1, Ratcliff, 1978.)

chi-square a very powerful test), the fits are actually quite good; certainly the convolution captures the overall shape of the distribution. Problems associated with truncation and outliers are discussed in the section entitled Moments and Cumulants.

Probability Mixtures of Distributions and Bimodality

Occasionally a model is developed that predicts bimodal reaction time distributions arising from a probability mixture of processes (e.g., Atkinson & Juola, 1973). The question arises as to whether Vincent averaging across bimodal distributions from individual subjects will produce a bimodal group distribution. In general, the answer is only under conditions in which the proportion of responses in each process is approximately the same across subjects. For example, Figure 5 shows the Vincent average cumulative distribution function for two distribution functions, each of which is a probability mixture of two processes. One distribution has a 25%–75% combination, and the other has a 75%–25% combination. The resulting group distribution is trimodal (i.e., has four points of inflection) and certainly does not reflect the bimodal nature of the individual distributions. In situations in which bimodality and probability mixtures of processes are expected, it is probably best to collect several hundred latencies per subject condition and investigate the individual latency distributions.

Moments and Cumulants

Moments have been used for many years to determine the shape of frequency curves either through skewness and kurtosis indices or by explicitly determining the frequency curve within Pearson's (cited in Elderton,

1906; Elderton & Johnson, 1969) system. Recently moments and cumulants have been used in the additive factor method for analysis of stage models (Sternberg, 1969; Sternberg, Note 2). In this section, three related problems in the use of moments and cumulants as sources of shape information are discussed. These problems are first that the variance associated with estimates of these measures is extremely large, second that the measures are very sensitive to outliers, and third that the measures give information about a part of the frequency curve that is of little theoretical interest.

To investigate the variability of moments and cumulants, expressions for moments and cumulants and their standard deviations must be derived. These expressions are derived for an explicit distribution function to allow estimation of numerical values. The convolution of normal and exponential distributions is chosen because it approximates the shape of reaction time distributions.

Moments are defined as follows (Kendall & Stuart, 1969):

$$\mu'_1 = \int_{-\infty}^{\infty} t f(t) dt;$$

$$\mu_i = \int_{-\infty}^{\infty} (t - \mu'_1)^i f(t) dt, \quad \text{for} \quad i > 1.$$

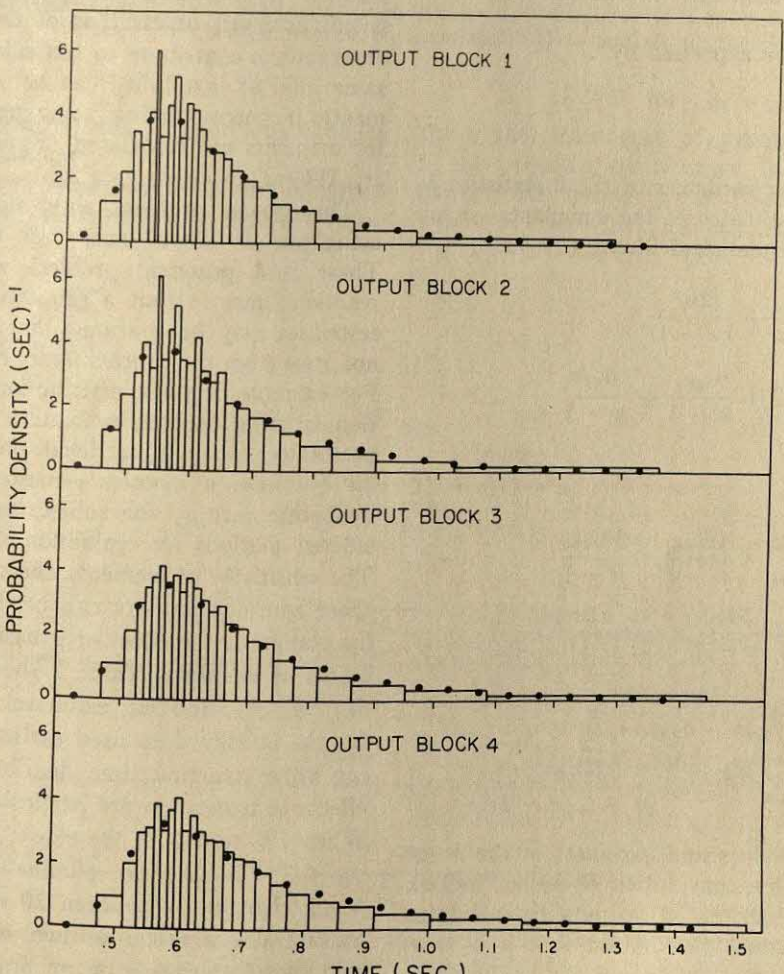

Figure 4. Group reaction time distributions for 2% quantiles for hits together with fits of the convolution model to the group distributions.

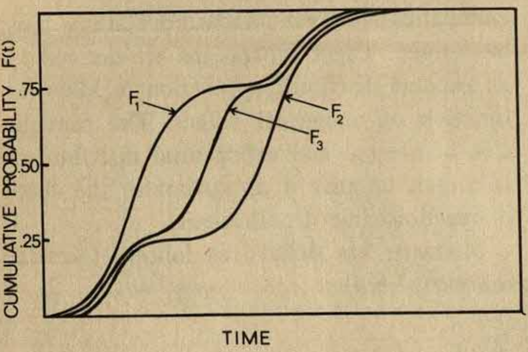

Figure 5. An example of group averaging of two bimodal distributions; $F_1(t)$ has 75% of the probability density in the first peak of the density function, $F_2(t)$ has 25% of the probability density in the first peak, and the resulting Vincent (1912) average distribution $F_3(t)$ is trimodal.

Cumulants are expressed by

$$\kappa_i = \mu_i, \quad \text{for} \quad i \leq 3;$$
$$\kappa_4 = \mu_4 - 3\mu_2^2.$$

The sampling variances of the k statistics k_i (unbiased estimates of the cumulants κ_i) are given by Kendall and Stuart (1969):

$$\text{var}\,(k_2) = \frac{\kappa_4}{n} + \frac{2\kappa_2^2}{(n-1)},$$

$$\text{var}\,(k_3) = \frac{\kappa_6}{n} + \frac{9\kappa_4\kappa_2}{n-1} + \frac{9\kappa_3^2}{n-1} + \frac{6n\kappa_2^3}{(n-1)(n-2)},$$

$$\text{var}\,(k_4) = \frac{\kappa_8}{n} + \frac{16\kappa_6\kappa_2}{n-1} + \frac{48\kappa_5\kappa_3}{n-1} + \frac{34\kappa_4^2}{n-1} + \frac{12n\kappa_4\kappa_2^2}{(n-1)(n-2)} + \frac{144n\kappa_3^2\kappa_2}{(n-1)(n-2)} + \frac{24n(n+1)\kappa_2^4}{(n-1)(n-2)(n-3)}.$$

Expected values and variances of the k statistics for the convolution of normal and exponential distribution can now be calculated. For the normal distribution, $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, and $\kappa_i = 0$, for $i > 2$; and for the exponential distribution, $\kappa_i = \tau^i(i-1)!$ To convolve two distributions cumulants are added; so for the convolution of normal and exponential distributions, $\kappa_1 = \mu + \tau$, $\kappa_2 = \sigma^2 + \tau^2$, $\kappa_3 = 2\tau^3$, and $\kappa_4 = 6\tau^4$. To estimate numerical values for cumulants and sampling variances of cumulants, values in the range of those found in Table 2 are used: $\mu = .5$ sec, $\sigma = .03$ sec, and $\tau = .2$ sec. Also, $\sigma^2 << \tau^2$, so that to an accuracy of 1% or 2% it is possible to neglect terms in σ compared with terms in τ. Now some numerical values can be calculated: For $n = 100$, $\kappa_2 = .040 \pm .011$, $\kappa_3 = .016 \pm .012$, and $\kappa_4 = .0096 \pm .0174$; for $n = 1{,}000$, $\kappa_2 = .040 \pm .004$, $\kappa_3 = .016 \pm .004$, and $\kappa_4 = .0096 \pm .0055$. From these values of cumulants and their estimated sampling standard errors, it can be seen that stability in the third and fourth cumulants is not achieved unless tens of thousands of observations contribute to the estimates. The same kind of instability can be seen in moments if corresponding sampling variances for moments are calculated (Kendall & Stuart, 1969).

The second problem with moments and cumulants is their sensitivity to outliers. There is a practical problem with outlier reaction times in that a proportion of these responses may be spurious, that is, they do not arise from the process under examination. For example, suppose distributional information is being used to evaluate a model that postulates a single retrieval process. Then an eyeblink, a moment's inattention, or a deliberate rest by the subject must be considered spurious for evaluation of the model. The sensitivity of moments and cumulants to these spurious outliers can be demonstrated by examining the effect of truncation on estimates of moments. Table 3 shows values of m'_1, m_2, m_3, and m_4, estimates of moments for the latency data used earlier (in obtaining group reaction time distributions). The effects of truncation are particularly striking. When 1% to 4% of the slower responses (2 sec $< T <$ 5 sec) are eliminated, mean latency changes by between 20 and 50 msec, variance by a factor of two, and the third and fourth moments by an order of magnitude. Thus, excluding outliers three or more standard deviations above the mean ($m'_1 +$

Table 3
Moments for Latency Data Truncated at 5 sec and 2 sec

Output block	$T < 5$ sec					$T < 2$ sec				
	m'_1	m_2	m_3	m_4	n	m'_1	m_2	m_3	m_4	n
					Hit					
1	6.70	6.98	.91	2.11	13,754	6.54	3.73	.18	.18	13,645
2	6.98	8.93	1.29	3.28	15,158	6.77	4.40	.21	.20	15,011
3	7.37	12.07	2.01	5.53	14,046	7.06	5.05	.25	.25	13,852
4	7.79	14.89	2.38	6.36	12,400	7.36	5.83	.28	.28	12,152
					Correct rejection					
1	7.31	9.76	1.44	3.59	13,885	7.08	4.75	.24	.23	13,722
2	7.59	11.54	1.69	4.15	15,118	7.29	5.37	.28	.27	14,890
3	7.98	15.36	2.42	6.13	15,050	7.53	6.03	.30	.29	14,709
4	8.24	20.26	3.73	10.49	13,587	7.63	6.49	.33	.33	13,192

Note. The values for m'_1 are in units of 10 msec, for m_2 in units of 10^4 msec^2, for m_3 in units of 10^8 msec^3, and for m_4 in units of 10^{11} msec^4; m'_1 is mean latency, m_2 is variance and m_3 and m_4 are the third and fourth moments, respectively.

$3(m_2)^{\frac{1}{2}} < 2$ sec) leads to enormous changes in higher moments.

The extreme sensitivity of higher moments to outliers is well-known, and Figure 6 illustrates the dependence of moments on tails of the frequency distribution. In Figure 6 is plotted the frequency distribution $f(x) = x^{-m}e^{-1/x}/[\Gamma(m-1)]$ for $m = 10.6$, together

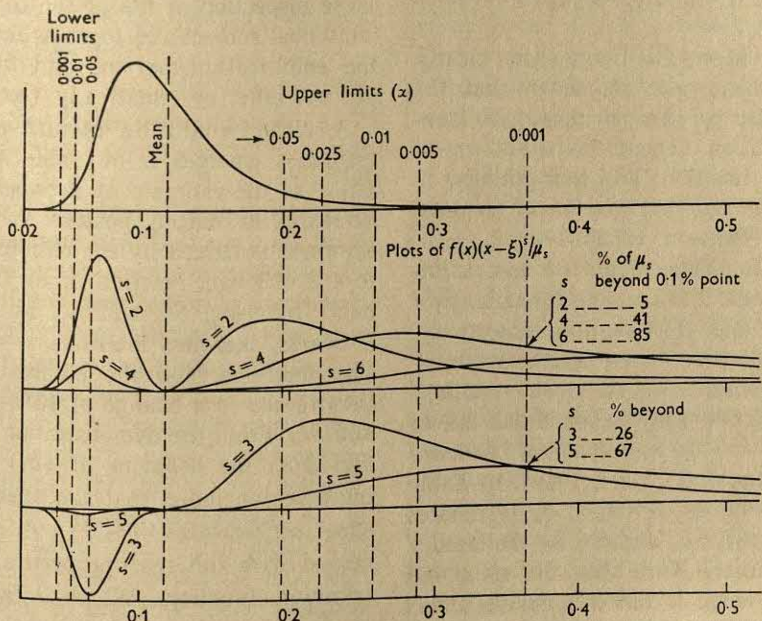

Figure 6. The distribution $f(x) = x^{-m} e^{-1/x}/[\Gamma(m-1)]$, for $m = 10.6$. The lower curves show $f(x) = (x-\xi)^s/\mu_s$, which are the normalized contributions to the moments μ_s, as a function of x, where ξ is the mean. (From "Some Problems Arising in Approximating to Probability Distributions, Using Moments" by E. S. Pearson, *Biometrika*, 1963, *50*, 95–112. Copyright 1963 by the Biometrika Trustees. Reprinted by permission.)

Table 4
Values of Skewness $(b_1)^{\frac{1}{2}}$, *Kurtosis* b_2, *and Pearson's*[a] *Measures of Skewness* $Sκ_o$ *and* $Sκ_e$ *for the Latency Data*

Output block	$T < 5$ sec				$T < 2$ sec			
	$(b_1)^{\frac{1}{2}}$	b_2	$Sκ_o$	$Sκ_e$	$(b_1)^{\frac{1}{2}}$	b_2	$Sκ_o$	$Sκ_e$
				Hit				
1	4.94	43.4	.447	.744	2.55	12.8	.533	.797
2	4.84	41.2	.475	.820	2.31	10.4	.577	.904
3	4.78	38.0	.478	.836	2.22	9.6	.596	.930
4	4.15	28.7	.492	.903	2.01	8.2	.605	1.008
				Correct rejection				
1	4.71	37.7	.483	.794	2.35	10.3	.587	.849
2	4.32	31.2	.495	.828	2.24	9.4	.600	.865
3	4.02	26.0	.503	.860	2.05	8.0	.623	.882
4	4.09	25.6	.480	.873	2.01	7.8	.608	.904

Note. $Sκ_o$ = *(mean − mode)/standard deviation*, where *mean* = m'_1 from Table 3, *standard deviation* = $(m_2)^{\frac{1}{2}}$ from Table 3, and *mode* is calculated from the convolution fits for $T < 2$ sec by setting the first derivative of the probability density function to zero and estimating $t\,[f'(t) = 0]$. $Sκ_e$ = 3 *(mean − median)/standard deviation.*
[a] Cited in Elderton (1906) and Elderton and Johnson (1969).

with the function $c(x) = (x - \xi)^s f(x)/\mu_s$ for $s = 2, 3, \ldots, 6$. Note that

$$\mu_s = \int_{-\infty}^{\infty} (x - \xi)^s f(x)dx,$$

where ξ is the mean. The figure shows clearly the third problem with moments—that the higher moments of an asymmetrical long-tailed distribution depend on the form of the frequency function (and thus outliers) in a region of the tail that may be of no practical interest (Pearson, 1963).

Third and fourth moments are used as indices of skewness and kurtosis through $(\beta_1)^{\frac{1}{2}} = \mu_3/(\mu_2)^{3/2}$ and $\beta_2 = \mu_4/\mu_2^2$, respectively. Pearson has proposed these alternative measures of skewness: $Sκ_o$ = *(mean − mode)/standard deviation* and, to avoid the use of the mode, $Sκ_e$ = 3*(mean − median)/standard deviation* (Kendall & Stuart, 1969). In Table 4 are shown values of $(b_1)^{\frac{1}{2}}$, b_2 (estimates of $(\beta_1)^{\frac{1}{2}}$ and β_2), $Sκ_o$, and $Sκ_e$ for the latency data used earlier. Note that the estimated value of the mode is rather unstable unless a fitted probability density function can be used to locate the mode (Elderton & Johnson, 1969). Thus, the mode used in the calculation of $Sκ_o$ was obtained from the convolution fit to the group data (for $T < 2$ sec) by setting the first derivative of the probability density function to zero. The truncated distribution ($T < 2$ sec) was chosen because inspection of fits of the convolution to individual subject's histograms indicated that the empirical histograms and fitted models did not differ systematically (see Figure 3).

A rather confusing picture of skewness estimates emerges from Table 4. By using $(b_1)^{\frac{1}{2}}$ as the estimate of skewness, skewness decreases as output position increases, and skewness is halved by the elimination of 1% to 4% of longer reaction times. On the other hand, by using $Sκ_o$ and $Sκ_e$ as measures of skewness, skewness increases as output position increases, and the elimination of outliers results in a change of 10%–20% in $Sκ_o$ and $Sκ_e$. From the demonstration in Figure 6 and from the behavior of $(b_1)^{\frac{1}{2}}$ and $Sκ$, it must be concluded that the alternative measures of skewness, $(b_1)^{\frac{1}{2}}$ and $Sκ$, are concerned with different properties of the distribution function. Which measure should be used depends on whether behavior of the central portion of the distribution function (indicated by $Sκ$) or behavior of the extreme tail of the distribution function [indicated by $(b_1)^{\frac{1}{2}}$] is of interest.

I attempted to fit Pearson's (cited in Elderton & Johnson, 1969) system of frequency curves using the moments in Table 3. The curve belongs to Pearson's Type VI class, but the system of frequency curves is not flexible enough to encompass the distributions used in Table 3. The start of a Type VI distribution is at some value, $a > 0$ (Elderton & Johnson, 1969). Calculating the value of a for one set of data in Table 3 gave a value of a around 6 sec, which is beyond the distribution cutoff value. Thus it seems that Pearson's system of frequency curves may not be as flexible as is generally thought (see Patel, Kapadia, & Owen, 1976, for a list of those distributions that belong to Pearson's system and those that do not).

To summarize, moments and cumulants higher than variance have little to offer as sources of shape information about reaction time distributions because of their extreme variability and because they provide information about the extreme tails of the distribution that is of little practical interest. More reasonable sources of shape information are mean, mode, median and standard deviation, together with Pearson's S_{K_0} and S_{K_e} measures of skewness.

A Further Alternative to Moments and Cumulants

Another way to obtain shape information from reaction time distributions is to fit an explicit distribution function and use the parameters of this distribution as a summary of shape. Ratcliff and Murdock (1976) have used the distribution resulting from the convolution of normal and exponential distributions (Equation 1) as an empirical summary of reaction time distributions in memory retrieval paradigms. For simple and choice reaction time paradigms, Snodgrass et al. (1967) have shown that distributions with a rounded mode and exponential tail (e.g., the gamma and so probably the convolution of normal and exponential distributions) are inadequate as descriptions of distribution shape. The distribution they find to give the best fits to their data is the double monomial distribution.

Presenting information about reaction time distributions by providing the parameters of an explicit distribution function (that fits adequately) has the great advantage that it is easy for anyone to reconstruct a distribution (from the formula) that has nearly the same shape as the raw data. This may prove extremely valuable for mathematical modelers who may not wish to invest a large amount of time in obtaining raw data until some initial checks have been carried out. Further examples and discussion of the use of explicit distribution functions as approximations to reaction time distributions can be found in Ratcliff (1978) and Ratcliff and Murdock (1976).

Conclusions and Summary

Information about reaction time distributions can prove very useful in model construction and model testing, but there are few methods available for analysis of distributions. In this article I have presented a method for obtaining group reaction time distributions from experiments in which there are as few as 10 observations per subject cell. The method essentially involves estimating latency quantiles for each subject and then averaging these over the group of subjects. Several distributions were shown to average to give another distribution of the same family with parameters that were the mean of the parameters of the individual member distributions. Several Monte Carlo studies were performed using the distribution that is the convolution of a normal and an exponential distribution, a distribution used to fit reaction time distributions. These studies showed that the parameters derived from the group distributions were the same as the parameters used to generate the individual pseudosubject distributions. Fits of the convolution model to group distributions derived from data combined from three large experiments gave parameters that were almost identical to average parameters from fits to the distributions of individual subjects. The close correspondence between these methods of estimating group averages shows that group distributions provide an excellent summary of

distributional information for the group and do not introduce any systematic bias into the estimate of shape.

Methods of deriving shape information that use moments and cumulants were evaluated, and three major problems were pointed out. First, estimates of the higher moments and cumulants have large standard deviations; for example, 10,000 observations may be needed before the standard deviation on the fourth cumulant is as low as 10% of the size of the fourth cumulant. Second, estimates of moments from data are extremely sensitive to outlier reaction times; the addition of 1% slow responses can change the fourth moment by a power of 10. This problem is particularly severe if an unknown proportion of the slow latencies are spurious, that is, if they are not a result of processes under examination. Third, Figure 6 shows that the third and fourth moments tell one about portions of the distribution that may be of no theoretical interest. It is suggested that the mean and standard deviation together with estimates of median, mode, and Pearson's skewness measures ($S\kappa_0$ and $S\kappa_e$) provide better information about distribution shape. These statistics are adequate, but may not be the most convenient statistics for conceptualizing the distribution or for fitting the distribution to more complex theoretical models. It is further argued that fitting adequate, explicit probability density functions to the observed reaction time distributions may provide more useful summaries of distributional information for researchers involved in mathematical modeling.

Reference Notes

1. Sternberg, S. *Evidence against self-terminating memory search from properties of RT distributions.* Paper presented at the meeting of the Psychonomic Society, St. Louis, Mo. November 1973.
2. Sternberg, S. *Estimating the distribution of additive reaction time components.* Paper presented at the meeting of the Psychometric Society, Niagara Falls, Canada, October 1964.

References

Atkinson, R. C., & Juola, J. F. Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance IV.* New York: Academic Press, 1973.

Bakan, D. A generalization of Sidman's results on group and individual functions, and a criterion. *Psychological Bulletin,* 1954, *51,* 63–64.

Box, G. E. P., & Muller, M. E. A note on the generation of random normal deviates. *Annals of Mathematical Statistics,* 1958, *29,* 610–613.

David, H. A. *Order statistics.* New York: Wiley, 1970.

Elderton, W. P. *Frequency curves and correlation.* London: Layton, 1906.

Elderton, W. P., & Johnson, N. L. *Systems of frequency curves.* New York: Cambridge University Press, 1969.

Estes, W. K. The problem of inference from curves based on group data. *Psychological Bulletin,* 1956, *53,* 134–140.

Green, D. M., & Luce, R. D. Detection of auditory signals presented at random times: III. *Perception & Psychophysics,* 1971, *9,* 257–268.

Hayes, K. J. The backward curve: A method for the study of learning. *Psychological Review,* 1953, *60,* 269–275.

Hohle, R. H. Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology,* 1965, *69,* 382–386.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 1, 2nd ed.). London: Griffin, 1969.

Kintsch, W. *The representation of meaning in memory.* Hillsdale, N.J.: Erlbaum, 1974.

Luce, R. D., & Green, D. M. A neural timing theory for response times and the psychophysics of intensity. *Psychological Review,* 1972, *79,* 14–57.

McGill, W. J. Stochastic latency mechanisms. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1). New York: Wiley, 1963.

Mosteller, F., & Tukey, J. W. *Data analysis and regression: A second course in statistics.* Reading, Mass.: Addison-Wesley, 1977.

Okada, R. Decision latencies in short-term recognition memory. *Journal of Experimental Psychology,* 1971, *90,* 27–32.

Patel, J. K., Kapadia, C. H., & Owen, D. B. *Handbook of statistical distributions.* New York: Dekker, 1976.

Pearson, E. S. Some problems arising in approximating to probability distributions, using moments. *Biometrika,* 1963, *50,* 95–112.

Pike, R. Response latency models for signal detection. *Psychological Review,* 1973, *80,* 53–68.

Ratcliff, R. A theory of memory retrieval. *Psychological Review,* 1978, *85,* 59–108.

Ratcliff, R., & Murdock, B. B., Jr. Retrieval processes in recognition memory. *Psychological Review,* 1976, *83,* 190–214.

Sarhan, A. E., & Greenberg, B. G. *Contributions to order statistics.* New York: Wiley, 1962.

Sidman, M. A note on functional relations obtained from group data. *Psychological Bulletin,* 1952, *49,* 353–374.

Snodgrass, J. G. Foreperiod effects in simple reaction time: Anticipation or expectancy. *Journal of Experimental Psychology Monograph,* 1969, *79*(3, Pt. 2).

Snodgrass, J. G., Luce, R. D., & Galanter, E. Some experiments on simple and choice reaction time. *Journal of Experimental Psychology,* 1967, *75,* 1–17.

Spence, K. W. *Behavior theory and conditioning.* New Haven, Conn.: Yale University Press, 1956.

Sternberg, S. Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist,* 1969, *57,* 421–457.

Vincent, S. B. The function of the viborissae in the behavior of the white rat. *Behavioral Monographs,* 1912, *1,* No. 5.

Received February 23, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 3, 462–492

Models of Jury Decision Making: A Critical Review

Steven Penrod and Reid Hastie
Harvard University

Several models of jury decision making are reviewed. In each instance the model is described and compared with related models, its assumptions are scrutinized, its fit to normative data is evaluated, and possible revisions and extensions of the model are discussed. Models reviewed include (a) multinomial decision schemes designed to adduce implicit decision rules used in jury decision making, (b) binomial models of jury voting that use simplifying assumptions about jury decision making to assess the impact of explicit decision rules and jury size on verdict distributions, (c) Bayesian models that use normative data to estimate prior probabilities of defendants' "convictability" and juror accuracy, (d) models that assess the relationships among jury size, decision rule, and jury accuracy, (e) models that examine the relationship between juror and jury errors, and (f) a computer simulation that uses simple assumptions about group persuasion and individual differences in jurors' resistance to persuasion to model results from empirical studies of jury decision making.

Until quite recently most of our knowledge about juror and jury behavior has been based on survey research such as Kalven and Zeisel's *The American Jury* (1966) and archival data collected in studies of jury utilization. Within the past decade these data have been supplemented by a growing body of experimental jury research (e.g., Boehm, 1968; Davis, Kerr, Atkin, Holt, & Meek, 1975; Landy & Aronson, 1969; Mitchell & Byrne, 1973; Saks, 1977; Simon, 1967; Simon & Kaplan, 1972; Strodtbeck, James, & Hawkins, 1957; Valenti & Downing, 1975; Padawer-Singer & Barton, Note 1); investigators have used simulated, laboratory juries to explore a wide variety of factors that theoretically affect juror and jury decision making. (For reviews of this research, see Davis, Bray, & Holt, 1977; Penrod, Note 2.)

This research was supported by a grant from the James Marshall Foundation and Grant BNS76-11321 from the National Science Foundation.

The authors would like to thank Dan Stefek for his mathematical and computer programming assistance and Charles Judd and David Kenny for their comments on the manuscript.

Requests for reprints should be sent to Steven Penrod, who is now at the Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706.

Even more recently a number of researchers have begun using survey and experimental data to develop mathematical and computer models of jury decision making, and these models are the focus of this review. The models can be classified into six categories: (a) *implicit-decision-rule* models that use mathematical techniques to determine the implicit rules that juries appear to use in deliberation, (b) simple binomial probability models that use the binomial expansion to mimic the effects of jury size and explicit decision rules on the distribution of jury verdicts, (c) Bayesian models that use Bayesian and binomial models to estimate the prior probability that defendants are guilty and the probability that jurors can accurately detect guilt or innocence, (d) binomial models of juror accuracy and juror satisfaction and their relationship to jury size and decision rules, (e) models of the relationship between juror and jury errors, and (f) a computer model that uses computer-simulated jurors and data from jury studies to explore the relationships among the initial distribution of individual juror opinions, individual differences in persuadability, rates of juror vote changes, coalition sizes within a jury, deliberation times, and the distribution of jury verdicts.

We examine these models giving particular attention to their structure, the plausibility of their underlying assumptions, their fit to available data, and their generalizability.

Implicit Decision Rules

Of the models reviewed here, the one that is most firmly grounded in empirical data—in the sense that it makes the fewest assumptions about jury behavior—is the model developed by Davis and his colleagues (Davis, 1973; Davis et al., 1975; Davis, Kerr, Sussman, & Rissman, 1974). Davis made a fundamental distinction between two types of decision rules that a jury uses: an explicit decision rule such as the requirement that juries reach a unanimous verdict and an implicit decision rule that describes the method by which a jury actually arrives at a verdict. The assumption, of course, is that juries may in fact operate under a decision rule different from the one given to them in a judge's instructions. To determine the implicit rule Davis has used two sources of data: Simon's (1967) study of the insanity defense, in which 30 12-person mock juries deliberated on a housebreaking case and 68 mock juries deliberated on an incest case, and the Davis et al. (1975) and Davis, Kerr, Stasser, Meek, and Holt (1977) studies of a rape case in which student jurors deliberated individually or in 6- or 12-person juries. These last juries were given either a unanimous or a two-thirds rule as their explicit decision rule.

In all of these studies the researchers asked the mock jurors to reveal their personally preferred verdict before deliberation began. Knowledge of the predeliberation distribution votes combined with knowledge of each jury's final verdict allows one to ask, can final verdicts be predicted by applying a particular decision rule to the initial distribution of individual juror votes? Davis (1973) called such a rule an "*implicit* social decision scheme (p. 99)." To determine the implicit decision rule Davis first examined the set of "distinguishable distributions of member preferences within a group (p. 101)." In general, these distributions can be determined by a multinomial expansion in which the set of initial distributions m equals

$$\binom{n+r-1}{r},$$

where n equals the number of mutually exclusive outcomes available to each group member and r equals the number of members. For jurors there are, of course, two possible outcomes (guilt or innocence), so $n = 2$. (When $n = 2$ the appropriate expansion is binomial, and the model is much simpler. A detailed discussion of the binomial expansion can be found in the next section of this article.) For 12-person juries ($r = 12$) there are 13 distinguishable initial distributions:

$$\binom{2+12-1}{12} = 13 = m.$$

These distributions range from 12 votes for guilt and none for innocence (12, 0) to 12 votes for innocence and no votes for guilt (0, 12). Similarly, in a 6-person jury there are 7 possible initial distributions, ranging from (6, 0) to (0, 6).

The question Davis has posed is, given actual data on the frequency of each initial distribution of individual juror verdicts, is there one general decision scheme that will accurately predict the distribution of final jury verdicts? Davis represented various possible decision schemes in the form of an $m \times s$ stochastic matrix in which s corresponds to the available group outcomes (e.g., a verdict of guilt, acquittal, or a hung jury). The entries in the matrix represent the probability that each of the m possible initial distributions will result in one of the s possible group outcomes. Each unique matrix thus represents a distinct decision scheme that may correspond to a jury's implicit decision rule.

Davis et al. (1975) have tested 13 different decision schemes that can be applied to the initial vote distributions for 12-person juries in which final verdicts can be of guilt, innocence, or a hung jury, and Davis, Kerr, Stasser, Meek, and Holt (1977) have tested 15 similar models for 6-member juries. Fortunately, the 3 decision schemes that have provided the best fits can be succinctly characterized verbally. For instance, Decision Scheme 8 from Davis et al. (1975) specifies that if two thirds or more of the jurors agree on a verdict on the initial ballot, this agreement determines the verdict. If two-thirds agreement is not obtained on the first ballot, the decision rule specifies that the jury will hang. For Scheme 7, majorities win,

Table 1
Distribution of Votes for Acquittal on First Ballot and Jury Decisions

	No. of votes for acquittal on first ballot											
	0		1–5		6		7–11		12		Total	% of
Final verdict	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	total
Not guilty	0	0	5	5	5	50	37	91	26	100	73	32
Guilty	43	100	90	86	5	50	1	2	0	0	139	62
Hung	0	0	10	9	0	0	3	7	0	0	13	6
Total *n*	43		105		10		41		26		225	
% of total		19		47		4		18		12		100

Note. These data are from Kalven and Zeisel (1966).

otherwise the jury is hung. Scheme 3 reflects a majority persuasion effect in which *persuasion* depends on the size of the initial majority: When 11 or 12 jurors agree (or 5 or 6 jurors agree in a 6-member jury), the verdict is determined; for distributions between (10, 2) and (6, 6) or (4, 2) and (3, 3) in 6-member juries, the juries yield guilty verdicts with probability r_g/r (where r_g is the number of jurors voting for guilt) and not-guilty or hung verdicts with probability $\frac{1}{2}[1 - (r_g/r)]$; and the distributions (5, 7) to (2, 10) [(2, 4) in six-member juries] yield not-guilty verdicts with probability r_{ng}/r and guilty or hung verdicts with probability $\frac{1}{2}[1 - (r_{ng}/r)]$.

Davis (1973) tested 5 different schemes on Simon's (1967) data and found that Scheme 3 made the best predictions for both of Simon's cases. Similarly, he tested 13 decision schemes on his own (Davis et al., 1975) data. The best fitting scheme overall was the two-thirds-majority model (Scheme 8), although other models provided better fits for particular jury sizes and explicit decision rules.

Finally, 15 schemes were tested on the data from Davis, Kerr, Stasser, Meek, and Holt (1977). The study used six-member juries that deliberated with an explicit 4/6 (i.e., four members out of six must agree) decision rule. For these juries a modified two-thirds rule—similar to Scheme 8 except that 75% of (3, 3) juries acquit and 25% hang—best fit the data. Davis et al. concluded that in general juries appear to be operating under an implicit "two-thirds, otherwise hung" decision rule even when the explicit, judge-instructed rule is unanimity.

Davis's approach has been applied by other researchers with similar results. For instance, Saks (1977) studied the effects of explicit decision rule and jury size on the distribution of jury verdicts in two experiments with a total of 85 juries. When Saks tested for implicit decision rules he also found support for Decision Schemes 3 and 8, but he found even stronger support for a power function rule suggested by Latane and Borden (Note 3).

Gelfand and Solomon (1975, 1977) have employed Davis's decision scheme method in their efforts to fit the data on 225 juries supplied by Kalven and Zeisel (1966). (Table 1 reproduces the Kalven and Zeisel data.) Although the first ballot distribution was not given for every possible initial jury split (nonunanimous majorities for guilt and innocence were separately pooled), with minor modifications in Davis's Scheme 3, Gelfand and Solomon estimated the probability of conviction to be .637, of acquittal to be .303, and of a hung jury to be .060—a very close fit to the Kalven and Zeisel data. The Gelfand and Solomon scheme, like Davis's Scheme 3, incorporates a strong majority persuasion effect, but elevates the probability of a hung jury for (10, 2) and (2, 10) initial distributions.

Although these results as a whole support the argument that juries may use something other than their assigned explicit rule, some caution is in order, for the studies reported to date have failed to demonstrate that one particular model consistently makes accurate predictions for all jury sizes and explicit decision rules. Indeed, as Davis et al. (1975) have noted, different implicit rules may apply to different

jury conditions; it may be that different implicit rules apply in criminal and civil cases, that different levels of a defendant's apparent guilt elicit different implicit rules, that complex cases involve yet another type of implicit rule, that different types of judge's instructions vary the implicit rule, and so forth. Only further research can resolve these problems.

An even more subtle and potentially more important methodological problem of the decision scheme analysis must be addressed. To date the evidence suggests that the rate of accurate prediction for individual trials is not reliably high. Most of the tests of implicit rules rely on a comparison of (a) the distribution of verdicts predicted from the application of the decision schemes to initial vote distributions with (b) the final distribution of actual verdicts. But without detailed inspection it is not clear that the final verdicts considered individually are consistent with the predictions made by the various decision schemes. Scheme 8, for instance, might predict an overall distribution of verdicts that would resemble the overall distribution of actual verdicts, but the entries in the stochastic matrix associated with the model might do a poor job of predicting actual verdicts on the basis of initial vote distributions. In other words, the model might fit the aggregated data, but for the wrong reasons. There could be an infinite number of matrices that would predict final distributions identical to those of Scheme 8, but only one of them would make the maximum number of correct predictions. Optimally, one would want to take a model and examine the accuracy of predictions for each entry in the matrix. In the case of Scheme 3, for instance, do 83% of the cases with an initial (10, 2) distribution of votes end up with a guilty verdict, 8% with a hung verdict, and 8% with an innocent verdict? Do 50% of the (6, 6) juries vote for guilt while the other 50% divide evenly to acquit or to hang?

The best fitting Scheme 8 (the two-thirds decision rule) is at least partially defective in this regard. One illustration of its failure is that it does not allow either for reversals of initial majorities (cases in which a minority on the initial ballot ultimately prevails) or for hung juries in cases in which eight or more jurors initially agree. And yet there is ample evidence (e.g., Padawer-Singer & Barton, Note 1) that initial majorities are reversed and that juries can hang even when eight jurors initially agree on a verdict. Some of the models allow for such reversals and provide more sources of hung juries, but unless the rules are fitted to actual outcomes on a jury-by-jury basis rather than on an aggregate distribution of outcomes, it is premature to say that one of the schemes accurately reflects the implicit decision rule.

In fact, the initial results obtained from nonaggregated analyses, although mixed, are on the whole disappointing. Grofman (1976) tested the fit of the two-thirds model on the Davis et al. (1975) and Davis, Kerr, Stasser, Meek, and Holt (1977) data and reported that the results were disappointing. We found that the two-thirds rule predicted the verdicts of 66 of 100 juries in the Kerr et al. (1976) study and 59 of 90 juries in the Davis, Kerr, Stasser, Meek, and Holt (1977) study. In fact, even the best fitting Scheme 15 (the modified two-thirds decision rule) mispredicts 10 of 90 juries. On the other hand, the standard two-thirds model predicted 24 of 25 juries in a study by Grofman and Hamilton (Note 4).

In a later section we discuss some of the factors that may contribute to the poor fit of social decision schemes when they are applied to individual juries, but we note briefly at this point that one major source of the problem may be a by-product of the case that Davis and his colleagues have used in their research. The case is one of rape, and it evokes different reactions from male and female subjects. In both Davis, Kerr, Stasser, Meek, and Holt (1977) and Kerr et al. (1976), approximately 60% of the females voted to convict on the first ballot, while only about 50% of the males voted to convict. During deliberation females changed their votes from conviction to acquittal at rather high rates (in the first study, 18.3% of the females shifted from guilt to innocence compared with a 5.5% shift in the other direction, and males shifted relatively little in either direction). As a result of the higher rate of change to acquittal by females, the overall distribution of verdicts is shifted in the direction of acquittal. Since the social decision schemes typically assume that jurors are equally likely to shift in either direction,

they fail to account for the bias in female juror behavior and are therefore unable to capture the shift to acquittals. Rape cases may be peculiarly susceptible to this sort of phenomenon, and this observation underscores the fact that case type has a significant impact on the deliberation process.

Probabilistic Models

Earlier it was noted that Davis's jury model is built around the binomial expansion. There are other modeling efforts that also make use of the binomial expansion to mathematically evaluate the effects that jury size and decision rule have on the distribution of jury verdicts. These issues are of practical significance and have been the subject of litigation before the U.S. Supreme Court. In 1970 the Court held that six-member juries did not violate a defendant's right to a trial by jury (*Williams* v. *Florida*[1]), and in 1972 the Court held by a narrow margin that nonunanimous juries were constitutional (*Apodaca* v. *Oregon*[2] and *Johnson* v. *Louisiana*[3]). More recently, in *Ballew* v. *Georgia*,[4] the Court, with Justice Blackmun writing the Court's main opinion, cited a number of empirical and theoretical studies to support the holding that juries with fewer than six members violate a defendant's right to a trial by jury.

The court has not yet reached the question of the constitutionality of nonunanimous six-member juries, but since the issue is likely to be pressed in the courts, empirical research and mathematical models directed to the question have acquired additional importance. The mathematical models that are most relevant are the binomial models of Walbert (1971) and Saks and Ostrom (1975). The binomal expansion is a mathematical expression that facilitates the calculation of the probability that a specified number of successes (or failures) will occur in a given number of trials (in the nonlegal sense), where trials are independent of one another and the probability of success is the same for each trial. More simply (using juries as an example), the binomial theorem applies in situations in which there are two possible outcomes (e.g., a juror can vote for either guilt or acquittal), in which there are a fixed number of trials (e.g., there are 12 jurors who must vote), in which outcomes have identical probabilities (e.g., jurors are randomly drawn from a large jury pool in which a certain percentage of jurors will vote for guilt), and in which trials are independent (i.e., a juror's initial judgments of guilt or innocence are independent).

To apply the binomial theorem to the jury it is clear that several assumptions have to be made. First, it must be assumed that all jurors are prepared to vote for either conviction or acquittal; there can be no undecided jurors. It is not clear how often this assumption is violated in practice, but for statistical purposes it must be assumed that all jurors have at least some inclination—however small—toward conviction or acquittal. Second, it must be assumed that jurors are drawn from a larger pool of jurors, in which a certain percentage of the jurors will vote for conviction or acquittal. Finally, it must be assumed that the jurors have not influenced one another's judgments prior to deliberation (this is probably a fairly reasonable assumption in light of the fact that jurors are typically cautioned not to discuss a case or to form an opinion until they have heard all the evidence).

With these assumptions it is possible to use the binomial theorem to determine the probability that a jury will have sufficient votes to convict a defendant on the first ballot:

$$p(G) = \sum_{i=Q}^{n} \binom{n}{i} g^i (1 - g)^{n-i}, \qquad (1)$$

where $p(G)$ is the probability that a jury will produce sufficient votes to convict on the first ballot, n is jury size, Q is the minimum number of votes required to convict (required quorum), g is the probability that a randomly selected juror will vote for guilt, and $1 - g$ is the probability that a randomly selected juror will vote to acquit.

Note that in the case of a unanimous decision rule, $p(G)$ can simply be written as $p(G) = p^n$. The same theorem can also easily be written to determine the probability that a

---

[1] Williams v. Florida, 90 S.Ct. 1893 (1970).
[2] Apodaca v. Oregon, 92 S.Ct. 1628 (1972).
[3] Johnson v Louisiana, 92 S.Ct. 1635 (1972).
[4] Ballew v. Georgia, 98 S.Ct. 1029 (1978).

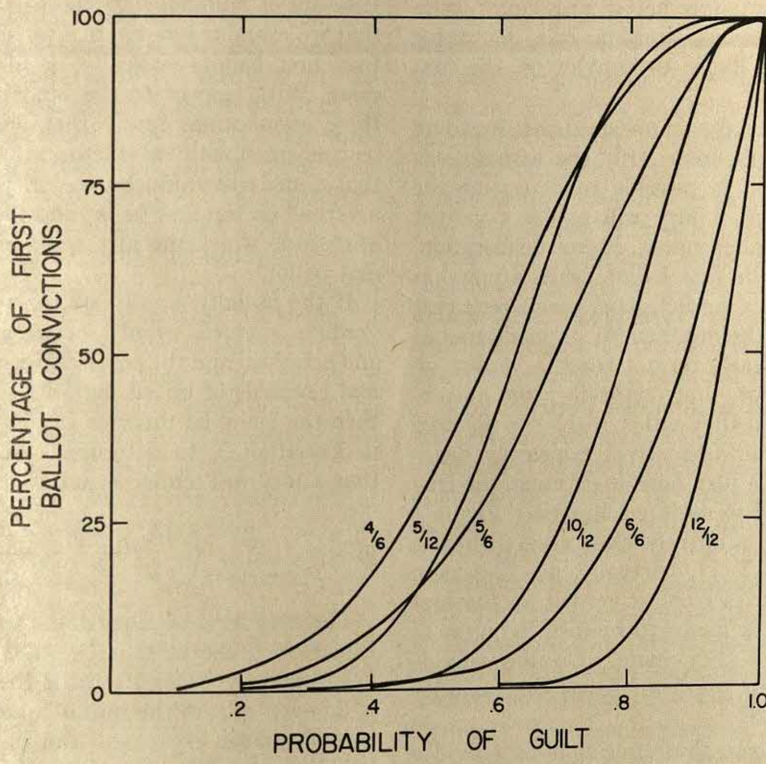

Figure 1. Percentage of first ballot convictions as a function of probability of guilt, according to a single, binomial decision rule model. (Denominator of parameters indicates jury size, and numerator indicates decision rule verdict requirement, i.e., the number of votes needed for a conviction.)

jury will produce sufficient votes to acquit on the first ballot:

$$p(I) = \sum_{i=Q}^{n} \binom{n}{i} g^{n-i}(1-g)^{i}. \qquad (2)$$

Thus, the probability $p(V)$ that a jury will have sufficient votes to produce either a guilt or an acquittal verdict on the first ballot is the sum of the probabilities from Equations 1 and 2:

$$p(V) = \sum_{i=Q}^{n} \binom{n}{i} g^{i}(1-g)^{n-i} + \sum_{i=Q}^{n} \binom{n}{i} g^{n-i}(1-g)^{i}.$$

One can, of course, vary jury size n, the size of the required quorum Q, and the probability that a randomly selected juror will vote for guilt g and use the binomial theorem to determine the probabilities that juries of varying size, drawn from differing jury pools, and who use different decision rules will produce first ballot verdicts for guilt or acquittal (one need only consult appropriate tables of the binomial distribution). Figure 1 summarizes the probabilities of first ballot verdicts for juries of 6 and 12 members who use two-thirds, five-sixths, and unanimous decision rules, where the probability of an individual juror voting for guilt ranges from 0 through 1.0. (Note that the figure also shows the probability of an acquittal—when $1 - g$ is substituted for g, the results are exactly symmetrical.)

Clearly, reductions in the size of the jury and relaxations of the quorum requirements produce similar effects: They heighten the probability that a jury will produce a verdict on the first ballot. Furthermore, if g is in any way an index of a defendant's objective guilt (irrespective of the evidence against the defendant) and juries can accurately detect such guilt on the basis of the evidence presented to them, then it is clear that objectively

guilty defendants fare better with larger juries using a unanimous decision rule, for these juries are less likely to convict on the first ballot.

Thus far we have emphasized that by using the binomial theorem with the assumptions made earlier, it is possible to determine the probability that a jury will render a verdict on the first ballot taken during deliberation. Events after the first ballot, when a jury has failed to reach a verdict, have been a recurrent problem for the mathematical modelers, as most juries clearly do not reach a verdict on the first ballot, but typically continue to deliberate until they either reach the required quorum or find themselves hopelessly deadlocked (when a jury finds itself hung, the trial judge is forced to declare a mistrial). Figure 1 shows that in relatively close cases (where g ranges from .3–.7), virtually no 12-person, unanimous juries reach a verdict on the first ballot. What is likely to happen in juries of this sort? Do they hang? Do they divide evenly in their verdicts? Does the majority tend to prevail?

One answer to these questions is provided by Walbert's (1971) analysis. As one has already seen, Davis's (1973) implicit-decision-rule analysis suggests that initial majorities tend to determine final verdicts; Walbert cited additional evidence to support the claim that deliberation after the first ballot is fundamentally irrelevant to the mathematical analysis. He argued as follows: (a) Small-groups research shows strong majority persuasion effects—with complex judgments minorities tend to conform to the judgments of the majority—that are accentuated by external pressures (such as judges' instructions) and are most evident in group discussions with leadership structures resembling those found in the jury room. (b) Empirical evidence indicates (Kalven & Zeisel, 1966) that majority persuasion operates in about 93% of all cases (minorities prevailed in 3%, and 4% ended with a hung jury). (c) Juries in which jurors initially divide evenly (6 to 6) tend to split evenly for conviction and acquittal (Kalven & Zeisel, 1966).

If one takes these data and assumptions as valid and further assumes that the reported discrepancies (i.e., reversals of initial majorities) are of minimal importance (assumptions that we examine below), it is possible to argue that first ballots basically decide almost all cases. With respect to the binomial theorem, these assumptions imply that one need only be concerned with two factors: the probability that a jury will produce a majority of votes for a verdict on the first ballot and the disposition of cases in which the jury splits evenly on the first ballot.

If the initially evenly split cases yield final verdicts divided evenly between conviction and acquittal and the possibilities of hung cases and reversals of initial majorities are ignored, then the binomial theorem can be modified, as in Equation 3, to determine the probability that a jury will render a guilty verdict:

$$p(G) = \sum_{i=(n/2+1)}^{n} \binom{n}{i} g^i + \frac{g^{n/2}(1-g)^{n/2}}{2}. \quad (3)$$

The first term of Equation 3 is identical to Equation 1, except that Q is the majority of jurors and the second term is the probability of an even split in the initial ballot divided by two. A similar expression can be constructed for the probability of an innocent verdict and is analogous to Equation 2:

$$p(I) = \sum_{i=(n/2+1)}^{r} \binom{n}{i} g^{n-i} + \frac{g^{n/2}(1-g)^{n/2}}{2}. \quad (4)$$

Note that by definition the sum of Equations 3 and 4 is 1.00. Furthermore, in these equations the decision rule is essentially fixed at *majority wins*, and only jury size affects the distribution of guilty and innocent verdicts.

Walbert's decision rule is quite similar to Davis et al. (1975) Scheme 7, except that Walbert eliminated hung juries and assigned evenly divided juries equally to guilt and acquittal. However, where Davis examined results by applying his schemes to initial votes from experimental studies to determine their fit to the data, Walbert made a more general argument. Starting from the assumptions outlined above, Walbert calculated the effect of jury size on the distribution of jury verdicts given a wide range of binomially distributed initial individual votes and using (as in Saks & Ostrom, 1975) varying levels of g (the probability that a randomly selected juror will vote for guilt). Walbert's results indicate that

a reduction in jury size raises the probability of conviction when $g < .5$: For example, when $g = .4$, the use of a 6-member jury increases the probability of conviction from .25 (for the 12-member jury) to .32. And when $g = .2$, the 6-member jury is six times more likely to convict (6% vs. 1%). Parallel results for acquittals are obtained when $g > .5$.

If one makes the additional assumption that g is a fairly reliable indicator of a defendant's "convictability" (i.e., g indexes the weight of the evidence against the defendant), then Walbert's analysis implies that the more convictable defendants are better off with the smaller juries, whereas relatively unconvictable defendants fare better with large juries.

Furthermore, if one regards convictions of relatively unconvictable defendants or acquittals of relatively convictable defendants as errors, it is clear that the larger jury is preferable. Indeed, it can be argued that the optimal pattern of conviction and acquittal would break sharply at $g = .5$. Juries would always convict when g was greater than .5 and acquit when g was less than .5. With these assumptions, optimality would be achieved with infinitely large juries using majority rule. Coincidentally, this optimality rule also maximizes the likelihood that jury verdicts will accurately reflect the attitudes and values of the community from which the jury pool has been drawn (the representativeness of the pool itself would also be a consideration). Indeed, the binomial theorem can be used to determine the probability that a specified number of jurors (analogous to Q) will be selected from a jury pool in which a specified trait (say, race or sex) occurs at a particular rate (analogous to g). Lempert (1975) discussed this representativeness question at length, and in his article he summarized the probabilities that two, one, or no jurors will be selected with the appropriate trait in 6- and 12-member juries for various rates of trait occurrence in the jury pool.

Given the assumptions made in Walbert's (1971) analysis, it is clear that the 6-member jury is inferior to the 12-member jury: It produces more errors and is less likely to represent the population from which it is drawn.

We note, in passing, that there is yet another way to think about g that could provide a quantifiable definition of *reasonable doubt*. If we were to define g_1 as the probability that a randomly selected juror will say that a defendant committed the charged acts beyond a reasonable doubt and g_2 as the probability that a randomly selected juror will vote to convict the defendant under a more-likely-than-not standard of proof, then we could determine a probabilistic value for RD (the additional certainty required for a reasonable-doubt standard) such that $RD = g_1 - g_2$.

If by empirical methods we found that RD was greater than zero, we could use Walbert's formulation of the binomial decision model to examine the effects of a relaxation in the standard of proof required in criminal cases. If, for example, $g_1 = .6$, $g_2 = .5$, and $RD = .1$, a relaxation from the reasonable-doubt standard to the more-likely-than-not standard would increase the probability of conviction from .5 to .75 in the 12-member jury and from .5 to .68 in the 6-member jury (see Walbert's Figure 1). Similar comparisons can be made for different values of g and RD. In every instance (for $RD > 0$) the relaxation of the standard accentuates the differences in 6- versus 12-member error rates.

To date the empirical evidence regarding the value of RD is mixed. In an experimental study involving a theft case, Cornish and Sealy (1973) found that the probability of a conviction under the reasonable-doubt standard was .50 compared with .51 for the more-likely-than-not standard. Simon and Mahan (1971) asked judges, jurors and, students to rate the chances out of 10 that a defendant committed a crime if convicted beyond a reasonable doubt and if convicted by a preponderance of the evidence. All three groups gave a mean rating of approximately 8.5 out of 10 for reasonable doubt; but whereas judges rated preponderance of the evidence at 5.5, the other two groups gave a rating of 7. Similarly, in Kerr et al. (1976), subjects individually judged a videotaped rape trial using one of three definitions of *beyond a reasonable doubt*. Of the jurors who formed a verdict preference after viewing the trial, 51% of those using a stringent definition (moral certainty), 61.2% of those with no definition, and 66.3% of those using a lax

definition (substantial doubt) voted to convict the defendant. When the subjects rated the probability of guilt associated with the standards, the mean rating for the stringent definition was .87 compared with .83 for the other two definitions. Nagel, Lamm, and Neef (Note 5), using a normative decision theory model, reported that student conviction probability thresholds averaged about .55 for various types of criminal cases. Thus, the preliminary evidence indicates that the standards of proof and the definitions of those standards can affect juror judgments, although perhaps not to the extent intended by the legal system. To date there is little research on the factors that may affect the value of RD: the quality of judges' instructions, differential experience with legal decision making, individual differences in jurors, differences in types of defendants and indictments, differences in available verdict alternatives, and differences in the types of judgments a jury must make (such as resolving issues of eyewitness identification vs. selecting a verdict or determining whether a defendant is guilty or innocent of any offense vs. determining which offense a defendant has committed).

Before leaving Walbert's (1971) model, it may be useful to note the points at which his analysis appears to be on the weakest ground. Unlike Davis, who attempted to account for reversals of initial majorities and hung juries in some of his models, Walbert assumed that these phenomena are unimportant when compared with the strong majority persuasion effects that he found in empirical data such as Kalven and Zeisel's (1966). However, as Gelfand and Solomon (1974) pointed out, it is not clear that Walbert has properly characterized the Kalven and Zeisel data (see Table 1). These data summarized the relationship between initial ballot distributions and final verdicts for 225 criminal cases tried in Chicago and Brooklyn. On close inspection, it turns out that for all cases in which the jury was not evenly divided (6 to 6), the initial majority ultimately prevailed in 91.5% of the cases (compared with Walbert's 93%), and the jury hung in 6.0% (compared with Walbert's 4%). Even assuming that Walbert's characterization of the Kalven and Zeisel data is the best possible, his model fails to account for more than 8% of the total cases (six reversals of majorities and 13 hung juries out of 225 cases), and it is not clear what implications these error cases have for his conclusions about the effects of jury size.

Beyond the question of the model's basic fit, one might want to know whether (and in what ways) cases involving reversals of majorities and hung juries differ from cases exhibiting majority persuasion effects. Intuitively, there is good reason to think that such cases are the most difficult for juries to decide; the evidence is not sufficiently compelling in favor of a guilty or an innocent verdict for jurors to be able to readily agree. In other words, these cases may tend to be the ones in which g is close to .5. Some evidence for the difficulty of such cases is provided by the results from Padawer-Singer and Barton's (Note 1) study of 92 6- and 12-member mock juries that deliberated after viewing a videotaped murder trial. This case was evidently fairly difficult to decide, for the probability that a juror randomly selected from the pool of jurors who viewed the trial would vote for guilt was $\approx .47$. And (summing across 6- and 12-member juries who used both unanimous and nonunanimous decision rules) for the 70 juries with initial majorities for a verdict, the majority prevailed in only 68.6% of the cases, the minority prevailed in 12.9%, and 18.6% were hung. Of the 22 juries that were initially evenly split (6 to 6 or 3 to 3), 54.5% ultimately returned innocent verdicts, 40.9% returned guilty verdicts, and 4.5% returned hung verdicts.

The Padawer-Singer and Barton data suggest that Walbert's model may provide the poorest fit for those cases in which accuracy and representativeness are most critical—the "hard to decide" cases. It would be premature, however, to conclude that the postulated lack of fit for cases with g near .5 undermines Walbert's basic arguments about the effects of jury size on jury accuracy and representativeness, for as long as majorities prevail more often than minorities, larger juries should be preferred to smaller juries.

Bayesian Models

One important question that a researcher might ask about juror performance relates to

the juror's ability to accurately determine a defendant's guilt or innocence. One would like to know how often, under the best of circumstances, jurors and juries accurately determine a defendant's objective guilt or innocence. But of course there is no reliable method of determining objective guilt and innocence. One can only ask how reliable jurors and juries are at the task of assessing guilt and innocence from the evidence presented to them at trial (and we would hesitate to argue that the quality of trial evidence is necessarily related to a defendant's objective guilt).

The models examined thus far do not attempt to evaluate the acuity of jurors, but assume that the best available index of a defendant's guilt is the proportion of jurors in a jury pool who, after hearing all the evidence and testimony, are prepared to vote for conviction. In this section we examine a jury model that uses a Bayesian analysis to determine both the prior probability that a defendant is convictable and the probability that jurors will correctly assess this convictability.

In a series of articles, Gelfand and Solomon (1973, 1974, 1975, 1977) have developed a model based on Poisson's (1837) application of probability theory to jury verdicts. Following Poisson, Gelfand and Solomon began their analysis by suggesting that with adequate data two important parameters can be estimated:

1. T is the probability that before trial an accused is convictable, that is, the proportion of defendants brought to trial who are convictable or the probability that the weight of the evidence will be against a randomly selected defendant.

2. M is the probability that a juror will not vote for the wrong verdict, the probability that a jury will correctly assess and vote with the weight of the evidence against a defendant. (For the purpose of modeling, Gelfand and Solomon assumed that T and M are independent and that M is a common value for all jurors.)

Gelfand and Solomon made it clear that when they wrote about convictability, they were really discussing the standards of indictment and the community standards for conviction that might prevail in a criminal justice system and were not using convictability (or more loosely, *guilt*) to mean objective guilt.

The specification of these two parameters allows for the construction of a mixed binomial expression similar to Equation 1 in which it is possible to determine $W_{n,i}$, the probability that a jury with n members will cast exactly i votes for acquittal on the first ballot; $Y_{n,i}$, the probability that a jury with n members will cast at most i votes for acquittal on the first ballot; $p_{n,i}$, the probability that the defendant is guilty given exactly i votes for acquittal on the first ballot; and $P_{n,i}$, the probability that the defendant is guilty given at most i votes for acquittal on the first ballot, where

$$Y_{n,i} = \binom{n}{i}[TM^{n-i}(1-M)^i + (1-T)M^i(1-M)^{n-i}]; \quad (5)$$

$$W_{n,i} = \sum_{j=0}^{i} Y_{n,j};$$

$$p_{n,i} = \binom{n}{i}TM^{n-i}(1-M)^i/Y_{n,i}; \quad (6)$$

$$P_{n,i} = T\sum_{j=0}^{i}\binom{n}{j}M^{n-j}(1-M)^j/W_{n,i}.$$

Note that Y is the sum of correct and incorrect votes for guilt and innocence and that W cumulates the probabilities of votes for innocence from 0 through i votes. The expression for p is simply the proportion of votes for guilt that are attributable to guilty defendants, whereas P cumulates the proportion of votes attributable to guilty defendants from 0 through i votes.

Gelfand and Solomon (1973) demonstrated that the probability of a verdict is independent of T and that the correctness of a verdict is independent of jury size but does depend on the size of the quorum required for conviction. (This seeming contradiction of Walbert's analysis arises from the differences in the ways that the two models specify *correctness*.)

Using Poisson's (1837) data on French civil and criminal trials of the period from 1825 through 1833, Gelfand and Solomon (1973) obtained estimates of T and M by determining the probability of conviction in criminal cases at times when the required quorum for jury verdicts was either 7 of 12 votes (i.e., $W_{12,5}$ for the years 1825–1830) or 8 of 12 votes ($W_{12,4}$ for the years 1831–1833). With the knowledge that

$Y_{12,5} = W_{12,5} - W_{12,4}$, it was possible for Gelfand and Solomon to estimate the following overall parameter values: $T = .7494$, $M = .6391$, $P_{12,5} = .9406$, and $p_{12,5} = .9943$.

In their second article, Gelfand and Solomon (1974) applied similar methods to Kalven and Zeisel's (1966) data on first ballots and final verdicts in the 225 criminal cases cited by Walbert (1971). (See Table 1 for complete data.) In this instance their estimates of T and M were based on the overall distribution of verdicts reported in Table 1 and the various estimates this distribution provides for different Ys. Thus, 43 of the 225 (19%) juries produced first ballot, unanimous votes for guilt, so that one reasonable estimate of $Y_{12,0}$ is .19. Similar estimates of other Ys produce the following simultaneous equations:

$$Y_{12,0} = .19;$$

$$\sum_{i=1}^{5} Y_{12,i} = .47;$$

$$Y_{12,6} = .04;$$

$$\sum_{i=7}^{11} Y_{12,i} = .18;$$

$$Y_{12,12} = .12.$$

In their 1974 article, Gelfand and Solomon used the method of moments approach to find solutions for T and M, whereas in their 1975 article they treated the first ballot results as independent observations from a five-cell multinomial distribution and employed minimum chi-square and maximum likelihood estimation procedures to determine values for T and M. In each instance they also evaluated a three-parameter model in which M_1 is the probability that a juror will vote for guilt given a guilty defendant, M_2 is the probability that a juror will vote for innocence given an innocent defendant, and T is defined as before. The results of the three estimation procedures were roughly similar. The values of M_1 and M_2 clustered around .9 and did not appear to be significantly different (jurors appeared to be equally accurate in detecting guilt and innocence), while the values of T clustered around .7.

As noted earlier, the Gelfand and Solomon model is analogous to Walbert's binomial model, but uses two parameters rather than one. Gelfand and Solomon's model also allows a very simple comparison of the probability of conviction by 6- and 12-member juries for various values of T and M. Assuming majority persuasion and an equal split for guilt and innocence in juries who are initially evenly divided (precisely the assumptions made by Walbert), Gelfand and Solomon set the probability of conviction by a 12-member jury as

$$t = \sum_{i=0}^{5} Y_{12,i} + \tfrac{1}{2} Y_{12,6}$$

and the probability of conviction by a six member jury as

$$s = \sum_{i=0}^{2} Y_{6,i} + \tfrac{1}{2} Y_{6,3}.$$

Table 2 compares the values of t and s for $T = .2, .4, .6,$ and $.8$ and $M = .2, .4, .6,$ and $.8$.

Gelfand and Solomon (1974) took the position (one they retreated from in later articles)

Table 2
Probability of Conviction by 12-Man Jury, tw(T, M), *and 6-Man Jury,* s(T, M), *for Values of* T *and* M

T	$tw(T, M)$	$s(T, M)$
	$M = .2$	
.2	.793	.765
.4	.598	.588
.6	.402	.412
.8	.207	.235
	$M = .4$	
.2	.652	.610
.4	.551	.537
.6	.449	.463
.8	.348	.390
	$M = .6$	
.2	.348	.390
.4	.449	.463
.6	.551	.537
.8	.652	.610
	$M = .8$	
.2	.207	.235
.4	.402	.412
.6	.598	.588
.8	.793	.765

Note. Based on Gelfand and Solomon (1974).

that the differences in performance between 6- and 12-member juries are negligible over the full range of values for T and M. In fact, it is clear from Table 2 that whenever jurors vote correctly more than half the time, 12-member juries perform better than 6-member juries (i.e., they vote correctly more often). This is, of course, the same conclusion reached by Walbert (1971). Indeed, when one considers that there are thousands of criminal trials each year, it is also obvious that the otherwise negligible differences in performance may yield quite significant practical consequences, affecting thousands of lives.

Of course, we noted earlier that the Walbert model is less than ideal insofar as it fails to account for reversals of initial majorities and hung juries. In an effort to incorporate these two violations of the simple majority persuasion model, Gelfand and Solomon (1975) proposed that the Kalven and Zeisel data can provide the basis for more refined estimates of M, T, p, and P. Table 1 shows that all the cases with initial unanimous (12 to 0) votes for guilt ultimately returned guilty verdicts, 86% of the cases with between 7 and 11 votes for guilt did so, and half the cases with evenly divided juries did so, whereas 2% of the cases with initial though nonunanimous majorities favoring acquittal were reversed, the minority ultimately prevailing. On the basis of these results Gelfand and Solomon suggested the following equation for the probability that a jury will convict:

$$P_c = Y_{12,0} + .86 \sum_{j=1}^{5} Y_{12,j} + .50 Y_{12,6} + .02 \sum_{j=7}^{11} Y_{12,j}, \quad (7)$$

where, for example, $Y_{12,0}$ is the proportion of juries who begin deliberation with no votes for acquittal. Similarly, they suggested the following equation for the probability of acquittal:

$$P_a = Y_{12,12} + .91 \sum_{j=7}^{11} Y_{12,j} + .50 Y_{12,6} + .05 \sum_{j=1}^{5} Y_{12,j}. \quad (8)$$

The probability of a hung jury is thus $P_h = 1 - P_a - P_c$. Of course, not all acquittals and convictions are correct, but it is possible to determine the probability or proportion of convictions and acquittals that are correct. Thus,

$$P_{g/c} = (p_{12,0}Y_{12,0} + .86 \sum_{j=1}^{5} p_{12,j}Y_{12,j} + .50 p_{12,6}Y_{12,6} + .02 \sum_{j=7}^{11} p_{12,j}Y_{12,j})/P_c;$$

$$P_{i/a} = [P_a - (p_{12,12}Y_{12,12} + .91 \sum_{j=7}^{11} p_{12,j}Y_{12,j} + .50 p_{12,6}Y_{12,6} + .05 \sum_{j=1}^{5} p_{12,j}Y_{12,j})]/P_a.$$

Table 3 compares the results of the three methods of estimating the probabilities of interest for 12-member juries using Walbert's simple majority persuasion model and the more refined equations (7 and 8) proposed by Gelfand and Solomon.

To assess the reasonableness of the refined model, Gelfand and Solomon (1975) computed the values of P_c, P_a, P_h, $P_{g/c}$, and $P_{i/a}$ for different values of M and T and compared the results with the distribution of verdicts from the 225 Kalven and Zeisel cases cited earlier and with the overall distribution of verdicts from the 3,576 trials used in the entire Kalven and Zeisel (1966) study. Again, the best fits were produced with $M = .9$ and $T = .7$.

Finally, as noted above, Gelfand and Solomon (1975, 1977) used their maximum likelihood estimates of M and T in combination with Davis's (1973) social-decision-scheme analysis and by slightly modifying Davis et al.'s (1975) Scheme 3 produced a very good fit to the Kalven and Zeisel data: $P_c = .637$, $P_a = .303$, $P_h = .060$, $P_{g/c} = .9779$, and $P_{i/a} = .9385$—results that compare quite favorably with the results reported in Table 3.

Juror Accuracy and Satisfaction

Grofman (1976, in press) has employed a general binomial model similar to Gelfand and Solomon's for two purposes: to examine the effect of applying several simplifying assumptions to his model and to examine the implica-

Table 3
Comparison of Walbert's (1971) and Gelfand and Solomon's (1975) Jury Decision Models for 12-Member Juries

Study	Probability of outcome: Conviction	Acquittal	Hung	Defendant guilty given that jury convicts	Defendant innocent given that jury acquits
	Minimum χ^2 estimate				
Walbert (1971)	.6588	.3412	.0	.9986	.9938
Gelfand & Solomon (1975)	.5843	.3433	.0724	.9882	.9092
	Maximum likelihood estimate				
Walbert (1971)	.6897	.3103	.0	.9997	.9984
Gelfand & Solomon (1975)	.6189	.3155	.0656	.9918	.9129
	Method of moments estimation				
Walbert (1971)	.6999	.3001	.0	.9998	.9993
Gelfand & Solomon (1975)	.6340	.3059	.0601	.9930	.9175

tions for decision rule preferences of jurors' tolerance for verdict errors. We examine these analyses briefly, starting with the analysis of the simplifying assumptions.

Juror Acuity Model

Following Gelfand and Solomon (1973), Grofman (1976) associated a binomial p with the probability that a randomly selected juror will correctly judge innocent defendants to be innocent (P_{II}) and guilty defendants to be guilty (P_{GG}), where P_G is the proportion of defendants who are guilty, P_I is the proportion of defendants who are innocent (by definition, $P_I = 1 - P_G$), P_C is the proportion of defendants convicted by juries, P_A is the proportion of defendants acquitted by juries, P_H is the proportion of defendants whose juries hang, and q is the number of votes required for a verdict (either guilt or innocence) and corresponds to a de facto decision rule (when q votes are not obtained, the jury is assumed to hang). In developing his model, Grofman assumed for simplicity that $P_{II} = P_{GG} = p$ (i.e., that jurors are equally good at correctly determining guilt or innocence).

In some respects Grofman's model is a more general version of the Gelfand and Solomon (1973, 1974, 1975, 1977) model; Grofman's p and P_G correspond to Gelfand and Solomon's T and M.

As the first step in the construction of his model, Grofman examined the probability that a majority of jurors ($q = m$) in a jury with an odd number of jurors (N) will reach a correct verdict:

$$P\text{ (correct verdict)} = \sum_{h=m}^{N} \binom{N}{h} p^h (1-p)^{N-h}. \quad (9)$$

This expression is simply Equation 1 rewritten with $Q = m$ and with p as the probability that a juror will vote correctly. The general implications of this model are that when $p > \frac{1}{2}$, increasing the size of the jury also increases the probability that a majority will reach a correct verdict (while lowering the probability that a verdict will be reached); when $p = \frac{1}{2}$, the probability that a jury will reach a correct verdict is $\frac{1}{2}$ and is independent of jury size; and when $p < \frac{1}{2}$, the larger the size of the jury, the less likely it is to reach a correct verdict.

Equation 9 can also be used to create expressions for P_C, P_A, and P_{II}:

$$P_C = \sum_{h=q}^{N} \left[\binom{N}{h} p^h (1-p)^{N-h} P_G + p^{N-h}(1-p)^h P_I \right], \quad (10)$$

$$P_A = \sum_{h=q}^{N}\left[\binom{N}{h}p^h(1-p)^{N-h}P_I + p^{N-h}(1-p)^h P_G\right], \quad (11)$$

$$P_H = \sum_{h=N-q+1}^{q-1}\left[\binom{N}{h}p^h(1-p)^{N-h}\right]. \quad (12)$$

Equation 12 is a corrected form of Grofman's expression. Note that the first terms of Equations 10 and 11 determine the probability that q or more jurors will correctly vote for guilt or innocence, whereas the second terms determine the probability that q or more jurors will incorrectly vote for guilt or innocence (cf. Equations 5 and 6 in Gelfand & Solomon's, 1973, model). The P_H equation determines the probability that less than q jurors will concur on either guilt or innocence (note that when the decision rule is that the majority wins, P_H is the probability of an even split in juries with an even number of jurors, and $P_H = .0$ with odd-numbered juries). Note that P_H is independent of P_G and P_I.

Table 4
Distribution of First Ballot Votes When the Probability of a Juror Not Erring Equals .88 and the Probability the Defendant Is Guilty Equals .69

No. of votes	Total probability	Correct part	Incorrect part
Votes to convict			
12	.1488	.1488	.0
11	.2435	.2435	.0
10	.1826	.1826	.0
9	.0830	.0830	.0
8	.0255	.0255	.0
7	.0056	.0055	.0001
Votes to hang			
6	.0013	.0009	.0004
Votes to acquit			
7	.0026	.0025	.0001
8	.0115	.0115	.0
9	.0373	.0373	.0
10	.0820	.0820	.0
11	.1094	.1094	.0
12	.0669	.0669	.0

Note. Adapted from Gelfand and Solomon (1975).

Table 4 (based on Gelfand & Solomon, 1975) displays the probabilities of vote distributions from (0, 12) through (12, 0) that obtain from $p = .88$ and $P_G = .69$ and provides some insight into the operation of Equations 10–12. Note that the table is based on a 12-person jury using a majority decision rule. If the decision rule were 8 of 12 votes, the hung portion of the table would consist of the sum of the probabilities of the (7, 5), (6, 6), and (5, 7) distributions.

Grofman (1976) observed that if jurors are very accurate in assessing guilt and innocence, then the values for P_A and p_C are approximated by the first terms of Equations 10 and 11. From Table 4 it is clear that for a 12-member jury with $p = .88$ and a majority decision rule, Grofman's observation is quite correct—deleting the "incorrect" portions of the P_A and P_C terms loses only .02% of the cases.

If one further assumes that the decision rule is unanimity, then the equations simplify even further:

$$P_C \approx p^N P_G, \quad (13)$$

$$P_A \approx p^N P_I, \quad (14)$$

$$P_H \approx 1 - p^N, \quad (15)$$

where

$$P_G \approx \frac{P_C}{P_C + P_A},$$

$$P_I \approx \frac{P_A}{P_C + P_A},$$

$$p \approx (P_C + P_A)^{(1/n)}.$$

Grofman correctly observed that with the appropriate normative data, it is relatively easy to estimate the parameters P_G, P_A, and p, but in making his estimates he assumed that the appropriate data are the distribution of final jury verdicts when, in fact, the appropriate data are the distribution of initial ballots. Grofman assumed that he could treat q as though it were an effective decision rule (similar to the analyses of Davis, 1973, and Walbert, 1971); however, the simplifying assumptions that led him from Equations 10, 11, and 12 to Equations 13, 14, and 15 used restrictions on q that effectively changed the decision rule from a majority decision to unanimity rule. As Saks and Ostrom (1975) demonstrated, the effect of such restrictions is

to lower the probability that a jury will produce a verdict on the first ballot. (As can be seen from Table 4, when $p = .88$ and $p_G = .69$, only slightly more than one fifth of the juries can be expected to produce a verdict on the first ballot with a decision rule of unanimity.)

It is, therefore, appropriate to estimate the parameters of the simplified model using data on first ballot distributions. The estimates obtained using the Kalven and Zeisel (1966) data are $P_G = .6232$, $P_I = .3768$, $p = .9062$, and $P_H = .6933$. That the estimates are roughly equivalent to those made by Gelfand and Solomon (1975) should not surprise one, since Grofman's simplifying assumptions allow a straightforward estimation of the same parameters. The estimates are cruder because the simplified equations are based on initial vote distributions at the tails of the binomial distribution—(12, 0) and (0, 12)—but the similarity of the estimates produced by the two procedures tends to confirm the validity of Grofman's simplifying assumptions.

One might take this analysis one step further and use the data on final verdicts to assess first ballot parameter values. If, for instance, one regards final verdicts as the best available index of defendants' convictability or "acquittability," then the Kalven and Zeisel (1966) data in Table 1 indicate that 139 (61.78%) of the 225 defendants were convictable, that 73 (32.44%) were acquittable, and that in 13 cases (5.7%) the guilt or innocence of the defendant was ambiguous. One can now ask, what is the probability that a convictable (acquittable) defendant will be convicted (acquitted) on the first ballot? The answer is that 43 of the 139 defendants ultimately convicted were convicted on the first ballot (30.94%), whereas 26 of the 73 defendants ultimately acquitted were acquitted on the first ballot (35.62%).

Thus, the probability that an individual juror will vote to convict a convictable defendant on the first ballot is $.3094^{1/12}$ or .9069, and the probability that an individual juror will vote to acquit an acquittable defendant on the first ballot is $.3562^{1/12}$ or .9176. These results are in close agreement with the Gelfand and Solomon and Grofman estimates and suggest that jurors may be somewhat more reluctant to vote to convict an apparently guilty defendant on the first ballot than they are to vote to acquit an apparently innocent defendant. As before, of course, this analysis tells one nothing about the deliberation process itself; it merely provides a crude method for assessing first ballot, individual juror accuracy.

Juror Tolerance for Errors

Following the analyses of Rae (1969), Taylor (1969), Curtis (1972), and Badger (1972), Grofman (1976) has used his basic binomial model (Equations 8, 9, and 10) to explore the decision rule implications of various levels of juror tolerance for erroneous jury decisions. He first specified a trade-off ratio R that reflects the number of guilty defendants a juror would be willing to set free to avoid the erroneous conviction of one innocent defendant. Two ratios can be constructed using R: $P_R = R/(R+1)$ is the relative weight attached to avoiding false convictions (avoiding a Type I error) that Grofman analyzes as the weight attached to assuring correct convictions and $1 - P_R = R/(R+1)$ is the relative weight attached to assuring correct convictions of guilty defendants (avoiding a Type II error) that Grofman analyzes as the relative weight attached to avoiding false convictions. By way of example, if a juror is willing to set five guilty defendants free to avoid one false conviction, then $P_R = 5(5+1) = 5/6$ and $1 - P_R = 1/6$. (For discussions of Type I and Type II errors in legal contexts, see also Feinberg, 1971; Friedman, 1972; Tribe, 1971.)

By weighting the probability of correct voting to reflect the trade-off ratio P_R for Type I and Type II errors, a new equation can be formed for the weighted probability that a quorum q will produce a correct verdict:

$$\sum_{h=0}^{q-1}\left[\binom{N}{h}p^{N-h}(1-p)^h P_I P_R\right] + \sum_{h=q}^{N}\left[\binom{N}{h}p^h(1-p)^{N-h}P_G(1-P_R)\right]. \tag{16}$$

The second term of Equation 16 corresponds to the first term of Equation 10 and is simply the weighted probability that q or more jurors will correctly vote for conviction. The first

term of Equation 16 is related to the first term of Equation 11, but in addition to including all the cases in which q or more jurors correctly vote for acquittal, the first term of Equation 16 also includes the cases in which the jury hangs (from $h = 0$ through $q - 1$). In essence, Grofman's analysis treats hung juries as correct acquittals.

Although Grofman again regarded q as the effective decision rule, his analysis is appropriately treated as an analysis of first ballots rather than of final verdicts. This means, as we have noted before, that the analysis fails to account for those cases in which final verdicts are not characterized by majority persuasion (approximately 5% of the cases reported in Kalven & Zeisel, 1966). Given the limitations on one's knowledge about the deliberation process, Grofman's treatment of q as an effective decision rule is not fatal and has the advantage of simplifying his analysis. The question of interest, of course, concerns what decision rule will maximize the value of Equation 16.

For fixed ns, Grofman determined the values of q that maximize the expression for given levels of p, P_G, and P_R. Of greatest interest are those conditions in which juror accuracy is greater than one half (in which case q should be equal to or greater than a majority m). Grofman noted in particular that when $P_G > 1/2$, the optimal decision rule approaches or equals $q = n$ as P_R approaches 1. Grofman concluded "that it does not require an infinite value of R to justify in normative terms a decision rule requiring unanimity to convict!" (p. 11).

In a similar fashion, Grofman constructed an index of juror satisfaction (D) that reflects the subjective ratio of disappointment a juror experiences when an apparently innocent defendant is convicted compared with the disappointment experienced when an apparently guilty defendant is set free. (Disappointment is *subjective* insofar as a juror can rarely know with certainty that a defendant is objectively guilty or innocent and can only judge on the basis of trial evidence.) The relative disappointment of a Type I error can be denoted by $P_D = D/(D + 1)$ and a Type II error by $1 - P_D = D/(D + 1)$. Thus, if a juror is nine times more disappointed by an erroneous conviction than by an erroneous acquittal, $D = 9$, $P_D = 9/(9 + 1) = .9$, and $1 - P_D = .1$.

Using the trade-off ratio for disappointment, the weighted probability that a randomly selected juror will agree with a verdict (i.e., the juror's perception of the correctness of other jurors' assessments of the evidence) is expressed by

$$\sum_{h=q-1}^{N-1} \binom{N-1}{h} p^{h+1}(1-p)^{N-h-1} P_G P_D$$
$$+ \sum_{h=q-1}^{N-1} \binom{N-1}{h} p^{h+1}(1-p)^{N-h-1} P_I (1 - P_D). \tag{17}$$

(This expression includes a correction of a typographical error in Grofman, 1976.)

The first term of the expression is the weighted probability that a juror will join a quorum that will correctly convict, while the second term is the weighted probability that a juror will join a quorum that will correctly acquit. Grofman concluded from his analysis that the value of Expression 17 is maximized when the decision rule corresponds directly to the juror's trade-off ratio P_D (for a related analysis, see Curtis, 1972). Thus, if a juror is as disappointed by seeing two guilty defendants acquitted as by seeing one innocent defendant convicted ($P_D = \frac{2}{3}$), that juror should prefer a two-thirds decision rule. Again, of course, we caution that this analysis is most appropriate for first ballot distribution and would require modification if it were extended to final verdicts.

Both the Gelfand and Solomon and the Grofman analyses of prior probabilities of guilt and juror accuracy are based on data from a wide range of cases. In some of those cases the evidence of guilt was probably overwhelming, in some cases the evidence of guilt was probably fairly weak, and in many cases the evidence was probably relatively balanced. Because the estimates of the parameters used in the models are based on a range of cases, they may not be directly applicable to any particular case. In general, the results suggest that approximately 70% of the defendants brought to trial (in the cases examined by Kalven and Zeisel) are convictable and that jurors are on the

average about 90% accurate in their assessments of defendants' convictability. This does not mean that any particular defendant has a 70% chance of conviction, nor does it mean that one can expect 10% of the jurors to err in their judgments of a particular defendant's guilt. The Gelfand and Solomon and Grofman analyses simply are not appropriate for analysis of individual cases. Indeed, it is possible that the convictability and juror accuracy parameters (assumed to be independent in the two models) are in fact interrelated and that juror errors are most likely to occur in cases in which the evidence against and in favor of a defendant is relatively balanced. In cases in which the evidence points unequivocally in one direction or another, it is probably less likely that jurors will err. Existing data on juror behavior are not adequate to the task of determining the relationship between apparent guilt and jury acuity, but the reasonableness of the assumption of independence should be kept in mind when evaluating the parameter estimates provided by the models we have examined.

The aspect of both Gelfand and Solomon's and Grofman's analyses that we find most disturbing is the implication that there are only two distinct types of defendants: defendants who clearly should not be convicted on the weight of the trial evidence ($\approx$30% of all defendants) and for whom the probability that a juror will erroneously vote to convict is $\approx$.1 and defendants who clearly should be convicted on the weight of the trial evidence ($\approx$70% of all defendants) and for whom the probability that a juror will correctly vote to convict is $\approx$.9 (see sample results reported in Table 3).

Although such a view may provide a reasonable characterization of trial evidence and of jurors' assessment of that evidence, we think that this analysis obscures the fact that the evidence presented at trials probably varies widely in the extent to which it indicates a defendant's guilt or innocence. In some cases the evidence may overwhelmingly and unmistakably point to guilt or innocence, and with such evidence we would not be surprised to see unanimous first ballot verdicts. In other instances the evidence may be very close (some of it pointing to innocence and some pointing to guilt), and in these trials we would not be surprised to find the jury dividing equally for conviction and acquittal on the first ballot.

Basically, we think it misleading to conceive of juror accuracy in Gelfand and Solomon's terms—juror accuracy is, after all, limited by the quality of the evidence presented at trial. Optimal juror performance can probably be attained only under circumstances such as those outlined in the discussion of Saks and Ostrom's (1975) and Walbert's (1971) binomial models: With very large juries, defendants should only be convicted when a majority (or some other critical proportion) of the jurors vote to convict after hearing all the trial evidence. We think it more realistic to assume that the weight of trial evidence (the extent to which the evidence would convince a juror of the defendant's guilt) varies widely across trials and that the best indication of the variability in trial evidence weight is the variability in jurors' first ballot votes for conviction and acquittal.

Judging from Kalven and Zeisel's (1966) data on first ballot votes (Table 1), it appears that the weight of evidence is bimodally distributed and that in approximately 70% of the cases the evidence points to conviction. Because Table 1 provides only summary data on the distribution of votes (i.e., juries produce unanimous verdicts for conviction in 19% of the cases, produce nonunanimous majorities for acquittal in 18% of the cases, and produce unanimous verdicts for acquittal in 12% of the cases), one cannot fix the actual distribution of first ballot votes (or the underlying distribution of evidentiary weight that produces the first ballot distribution of votes). Still, one can test some possible distributions of evidentiary weight across trials (in which such weights represent the probability that a randomly selected juror who has heard the evidence from a particular trial will vote to convict) for their fit to the Kalven and Zeisel first ballot vote distribution.

For example, the Gelfand and Solomon (1975) model that produces the distribution of votes shown in Table 4 assumes a distribution of evidentiary weight in which the probability that a juror will vote to convict is .88 for 69% of the cases and .12 for 31% of the cases. This very simple bimodal distribution of evidence produces a moderately good fit to the Kalven

and Zeisel data. We have tested several other types of evidentiary weight distributions (relatively flat but skewed, unimodal, and bimodal) and have assumed in each instance that the weights are distributed in probability intervals of .1. We have found that bimodal distributions of evidence such as the one in Figure 2 produce the best fits to the Kalven and Zeisel data. (We caution that this estimation enterprise is crude; it is subject to the limitations of the data, it is post hoc, and it ignores that the distribution of evidence is probably continuous rather than discrete. Still, the distribution in Figure 2 is psychologically plausible and, as we demonstrate, produces a good fit to the distribution in Table 1.)

To explain our method briefly, one can see that the distribution of weights in Figure 2 implies that in 13% of all cases jurors are expected to judge the evidence against a defendant as unmistakably indicating that the defendant is guilty. In these cases the probability that jurors will vote to convict is equal to 1.0—Each of these juries will return unanimous verdicts for conviction (the expected outcome for a binomial $p = 1.0$ and for $N = 12$). In 8% of the cases Figure 2 implies that all jurors will vote for acquittal ($p = .00$). Similarly, in 16% of the cases the (binomial) probability that a randomly selected juror will vote to convict equals .8.

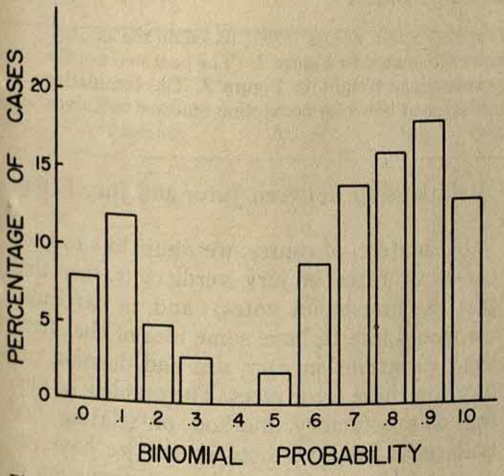

Figure 2. Hypothetical distribution, across cases, of evidentiary weight for conviction (based on data collected by Kalven & Zeisel, 1966).

By consulting binomial tables or using the binomial expansion for $N = 6$ and $N = 12$, one can determine the expected distribution of first ballot votes for the distribution of ps shown in Figure 2.

Figure 3 shows the expected distributions of votes for each of the binomial probabilities for 12-member juries. The figure also shows the cumulative probability of each vote distribution (i.e., if the evidence weight is distributed as shown in Figure 2, then in 12-member juries a unanimous verdict for acquittal should occur in 11.7% of the cases, with the bulk of these verdicts occurring in trials in which the evidence points unequivocally to innocence). The cumulative distribution of first ballot votes in the 12-member jury (shown at the top of Figure 3) provides a very good fit to the Kalven and Zeisel data (11.7%, 18.4%, 3.8%, 46.5%, and 19.4% in the model vs. 12%, 18%, 4%, 47%, and 19% in the respective categories for the normative data).

By making a few simple assumptions, one can also assess the distribution of first ballot votes for accuracy. For example, if one makes the crude assumption that juries should convict whenever the binomial probability of guilt is equal to or greater than .6 (i.e., when 60% or more of all jurors who might hear a case would vote to convict), should hang when $p = .5$, and should acquit when $p \leq .4$, then one can easily determine the number of juries that began deliberation with "errorful" first ballot distributions (e.g., in Figure 3 when $p \geq .6$, all the juries who have produced less than seven votes for conviction have made errors, since one has assumed that defendants with evidentiary weights of .6 or more should be convicted). In Figure 4 we display the proportions of cases in each category of initial votes, ranging from (0, 12) to (12, 0), that would convict, hang, and acquit if one applied the simple rules outlined above. Overall, the probability that a jury will begin deliberations with a majority erroneously preferring acquittal is 1.9% in 12-member juries, while the probability that a jury will begin deliberations with a majority erroneously preferring conviction is .1%. These values can be compared with the much lower error rates produced by Gelfand and Solomon's model (see Table 4). Although the absolute value of the estimates

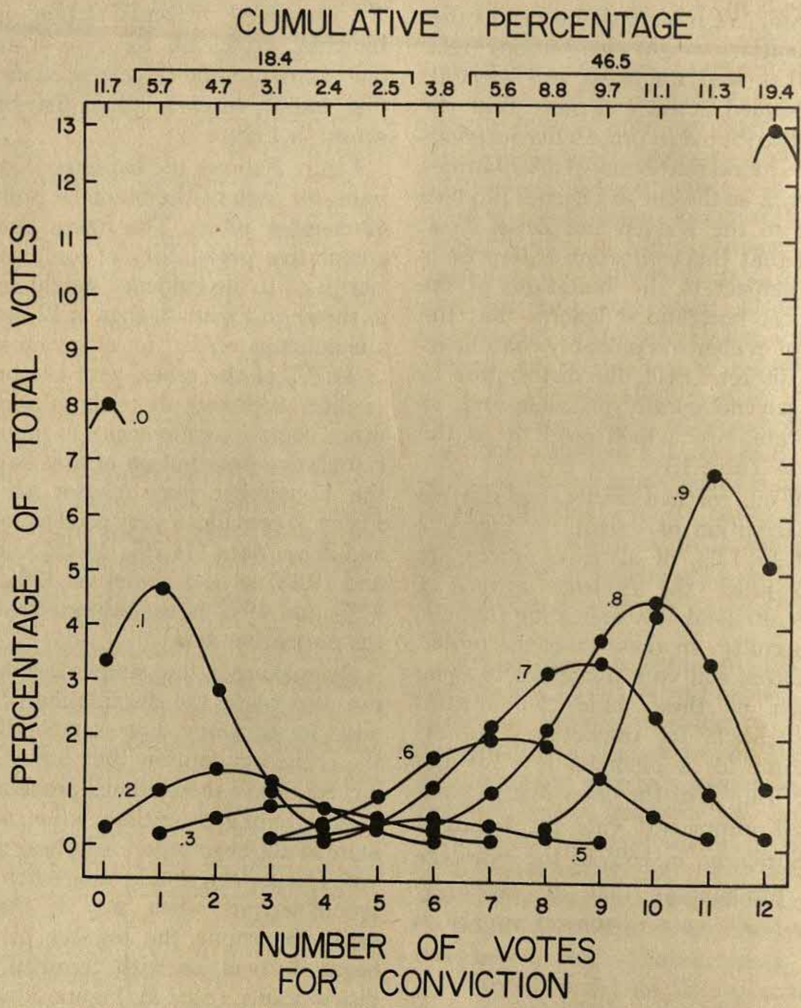

Figure 3. Distribution of first ballot vote percentages for conviction across juries, based on the assumption of a bimodal distribution of evidence such as the one illustrated in Figure 2. (The parameter of the plotted subdistributions in Figure 3 corresponds to evidentiary weight in Figure 2. The cumulative distribution totals, 18.4 and 46.5, correspond to the tabulation of votes for conviction reported in Kalven & Zeisel, 1966, and reproduced in Table 1.)

produced by the model we have developed in this section can be questioned (more complete data on the Kalven and Zeisel juries would heighten our confidence), we think that this model of juror accuracy is superior to Gelfand and Solomon's and Grofman's because it makes more plausible assumptions about the distribution of trial evidence weights. Since the distribution of first ballot votes produced by the model fits the normative data better than Gelfand and Solomon's model, we are also confident that Gelfand and Solomon's model underpredicts first ballot error rates.

Relationship Between Juror and Jury Errors

Ultimately, of course, we would like to know the error rates in jury verdicts (rather than just the first ballot votes); and, in particular, we would like to have some idea of the effects that variations in jury size and decision rule have on these error rates. The problem of finding a satisfactory method of relating first ballots to final verdicts is one we have encountered several times in our discussion of various mathematical models. Although we have criticized all of the proffered solutions to

this problem, we think that a decision scheme approach such as Davis's (1973) offers the most promise. Simply for purposes of illustration, we have applied the decision schemes in Table 5 to the initial vote distributions in Figure 2 to acquire a sense of the impact that different decision rules (unanimous vs. two thirds) might have on jury accuracy.

The decision schemes in Table 5 are based largely on the results of a computer simulation of jury decision making that is discussed later in this article (see also Penrod & Hastie, Note 6). These decision schemes are constructed to produce a slight postdeliberation bias toward acquittal, such as is observed in the Kalven and Zeisel (1966) data (in Table 1, more majorities reverse in the direction of acquittal than in the direction of conviction). These decision schemes are also roughly comparable with Gelfand and Solomon's (1975, 1977) version of Davis et al.'s (1975) Scheme 3.

Before discussing the results of this decision scheme analysis, it should be noted that these decision schemes reflect relatively unfavorable views of the deliberation process insofar as they assume that both correct and errorful initial majorities are equally likely to be reversed during deliberation. In other words, these decision schemes do not assume that deliberation serves to correct the errors made in first ballot votes. If deliberations did serve a correcting function, then one would expect the probability of a reversal in a first-ballot-error case to be higher than the probability of a reversal in an initially correct case. As one will see, our analysis indicates that the selection of optimal jury sizes and optimal decision rules

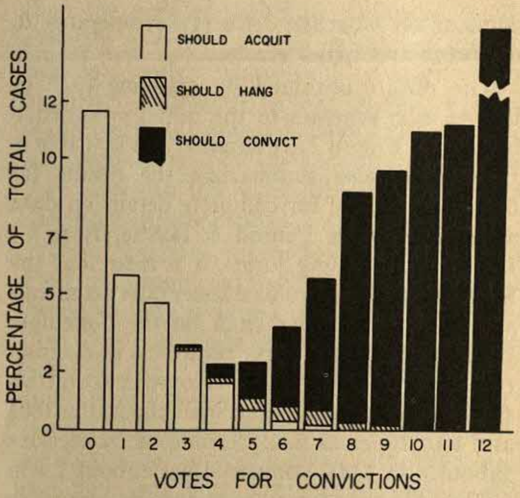

Figure 4. Distribution of percentage of votes for conviction across cases generated by the assumptions of the analysis outlined in Figures 2 and 3 and in the text. (Solid areas represent juries under conditions with evidentiary weights for conviction greater than .50, hatched areas represent juries with weights equal to .50, and clear areas represent juries with weights less than .50.)

Table 5
Decision Schemes

Votes for conviction	p of verdict in unanimity (12/12) rule			p of verdict in 8/12 rule		
	Acquit	Hang	Convict	Acquit	Hang	Convict
0	1.0			1.0		
1	1.0			1.0		
2	1.0			1.0		
3	1.0			1.0		
4	.92	.07	.01	1.0		
5	.70	.26	.04	.70	.26	.04
6	.35	.40	.25	.35	.40	.25
7	.14	.26	.60	.14	.26	.60
8	.04	.07	.89			1.0
9			1.0			1.0
10			1.0			1.0
11			1.0			1.0
12			1.0			1.0

Note. In 12/12 and 8/12 rule, the denominator indicates jury size and the numerator denotes the number of votes needed for a decision.

depends on whether deliberation operates to minimize first ballot errors.

The results obtained by applying each of the decision schemes to the initial vote distributions in Figure 2 are summarized in Table 6. (The table also summarizes the results for 6-member juries; for complete details on data and methods see Penrod & Hastie, Note 6.) If one ignores hung juries, it is clear that the 12-member juries produce fewer false acquittals and false convictions than do the 6-member juries: The smaller jury, regardless of decision rule, produces 50% more erroneous acquittals (about 5 in 100 compared with about 3 in 100) and four times as many erroneous convictions (about 6 in 1,000 compared with about 1.4 in 1,000). These same relationships are also evident in the probabilities that a defendant is guilty if convicted ($P_{G/C}$), innocent if convicted ($P_{I/C}$), innocent if acquitted ($P_{I/A}$), and guilty if acquitted ($P_{G/Z}$). (These probabilities can be compared with the 12-member probabilities produced by Gelfand & Solomon's model; see Table 3.)

Somewhat surprisingly, the error rates in nonunanimous juries are lower than the rates in the corresponding unanimous juries. The reason for this is that the decision schemes used to generate the final distributions of verdicts do not use the corrective factor discussed earlier. This point can be illustrated by examining the distribution of error cases in 12-member juries who produce eight votes for conviction on the first ballot. One can see in Figure 4 that most such juries' votes are consistent with the evidence, but a few "should" have begun deliberation with even splits. If these juries use a unanimous decision rule, one can see from Table 5 that 4% of the juries can be expected to reverse the original majority and produce verdicts for acquittals. Since the decision scheme does not distinguish between error cases and correct cases, the result is that 4% of the cases that start with errors will be corrected, while 4% of the cases that begin deliberation with majorities correctly preferring conviction will erroneously acquit. In essence, the decision scheme ultimately creates more errors than it corrects. For juries using an 8/12 decision rule, this problem is avoided—none of the original error

Table 6
Decision-Scheme-Generated Verdicts and Verdict Accuracy for 6- and 12-Member Juries Assigned Unanimous and Nonunanimous Decision Rules for N = 100 Cases

Jury decision	Acquittal			Hung case			Conviction			Total			Condition[a]			
rule[a]	COR	SHH	SHC	COR	SHC	SHA	COR	SHH	SHA	Convicted	Hung	Acquitted	$P_{G/C}$	$P_{I/C}$	$P_{I/A}$	$P_{G/A}$
12/12	27.23	.80	3.63	.42	3.51	.46	63.01	.64	.14	63.79	4.39	31.66	.9878	.0021	.8601	.1147
8/12	27.44	.82	3.27	.40	2.87	.45	64.16	.67	.14	64.96	3.72	31.53	.9877	.0021	.8703	.1037
6/6	26.76	.89	5.29	.31	3.74	.70	60.97	.81	.60	62.38	4.75	32.94	.9774	.0096	.8124	.1606
4/6	27.02	.91	4.95	.24	2.68	.48	62.38	.85	.60	63.83	3.40	32.88	.9773	.0094	.8218	.1505

Note. COR = correct, SHH = should have hung, SHC = should have convicted, and SHA = should have acquitted. $P_{G/C}$ = the probability that a defendant is guilty given that the defendant is convicted. $P_{I/C}$ = the probability that a defendant is innocent given that the defendant is convicted. $P_{I/A}$ = the probability that a defendant is innocent given that the defendant is acquitted. $P_{G/A}$ = the probability that a defendant is guilty given that the defendant is acquitted.
[a] The denominator of the decision rule indicates jury size, and the numerator indicates the number of votes needed for a decision.

cases are corrected, but no new error cases are created.

Although we have only limited confidence in the absolute value of the entries in Table 6 (since neither the initial distribution of evidence nor the decision schemes are well-grounded empirically), these results suggest that one's ability to distinguish among optimal decision rules may ultimately depend on one's ability to specify the extent to which jury deliberations serve to correct the "sampling errors" in the distribution of initial votes. Our guess is that a unanimous decision ultimately operates to minimize jury errors: First, it minimizes the probability of an incorrect verdict on the first ballot, and second, it maximizes the opportunities for jurors who have correctly assessed the weight of the evidence to communicate the grounds for their assessments to other jurors, who may then be persuaded to adopt a more accurate view of the evidence. Stated another way, we contend that it is unlikely that reversals of majorities are equally likely for initially correct and incorrect juries. Given the small magnitude of the advantages enjoyed by the non-unanimous juries under the equally likely decision schemes we have used (Table 6 shows that the advantages in the critical false-acquittal and false-conviction categories are very small), it is clear that even a relatively small correcting factor in jury deliberations would give the edge to the unanimous decision rules. As a practical matter, our results indicate that it is important to direct future research to the question of whether deliberation serves to minimize verdict errors. If there is something about the deliberation process that does serve to raise the probability that initial errors will be corrected, then the unanimous decision rule is probably preferable.

To conclude the section on models of juror accuracy we examine a model developed by Nagel and Neef (1975). Nagel and Neef have also tackled the question of the relationship among jury size, decision rules, and error rates, using an approach that is similar to but less general than Grofman's (1976) approach. Perhaps the most important difference between the two lines of analysis is that Nagel and Neef were concerned with defendants' objective or true guilt and innocence rather than the guilt or innocence indicated by the evidence presented at trial (what Gelfand and Solomon called "convictability"). A second important difference is that Nagel and Neef made specific assumptions about the values of the parameters they used in their model and then examined the results obtained with these specific parameter values.

Their analysis proceeded in two stages: First, they analyzed *individual* juror behavior (what we regard as *first ballot* behavior), and then they examined individual plus *collective* behavior (what we would consider to be an analysis of jury verdicts). Nagel and Neef's analysis can most clearly be understood by noting the assumptions about objective guilt and innocence and juror accuracy on which their analysis is based. First, they assumed that the number of truly guilty defendants brought to trial is relatively large (95%), while the number of truly innocent defendants is relatively small (5%). Although we consider the grounds for this assumption rather curious (it is based on an analogy to the .05 significance level used in statistical analysis), we do not quarrel with the reasonableness of the assumption; indeed, we have heard experienced defense attorneys make even lower estimates of the error rate in indictments. (Of course, since only about 10% of all criminal indictments reach trial—most defendants plea bargain—it is possible that truly innocent defendants are overrepresented at the trial stage.) The basic implication of Nagel and Neef's assumption is that it is relatively uncommon to find instances in which evidence sufficient for indictment points to the wrong defendant and that prosecutors are reasonably accurate in their indictments.

Next, Nagel and Neef assumed that 40% of all truly innocent defendants are erroneously convicted by juries (i.e., 2 in 100 cases yield false convictions), compared with a 70% conviction rate for truly guilty defendants (they made this estimate by reference to the Kalven & Zeisel, 1966, data that are presented in Table 1). Disregarding deliberation effects (a point to which we return), Nagel and Neef used the 40% and 70% figures to posit that the probability that an individual juror will erroneously vote to convict a truly innocent defendant is $.4^{1/12}$ or .926, whereas the proba-

bility that an individual juror will correctly vote to convict a truly guilty defendant is $.7^{1/12}$ or .971. Finally, Nagel and Neef assumed they could use these individual probabilities in the binomial expansion to determine the weighted probabilities of false convictions and false acquittals for various jury sizes and decision rules (fundamentally the method used in Grofman's, 1976, analysis). For ease of presentation, Nagel and Neef reported their results in terms of 1,000 cases in which 95% (950) of the defendants were truly guilty and 5% (50) were truly innocent. The results of their analysis (in which false convictions were given a weight of 10 compared with a weight of 1 for false acquittals) indicate that with binomial probabilities of .926 and .971, seven-member unanimous juries produce the minimum weighted sum of errors and 11/12 and 10/12 juries produce the lowest weighted sum of errors for nonunanimous juries of various sizes.

Unfortunately, we regard this analysis as faulty at several points. First, by taking the 12th root of the .4 and .7 probabilities, Nagel and Neef implicitly adopted the 12-member unanimous jury as the absolute standard against which all other jury sizes and decision rules are to be evaluated. Nagel and Neef failed to justify using $p^{1/12}$ as a standard individual probability. Second, because they used these probabilities in the binomial expansion to determine the probability that X or more jurors (where X is the decision rule quorum) in a jury with Y members will vote to convict in Y binomial trials, their initial analysis produced the curious result that all juries failing to attain the required quorum in Y binomial trials were treated as acquittals, even though they may have been only one vote short of the necessary quorum for conviction. As one saw before (e.g., in Walbert's, 1971, analysis), for any fixed binomial probability, any reduction in the number of binomial trials (i.e., any reduction in the number of jurors who constitute a jury) increases the probability that all the jurors (or some set proportion of the jurors) will agree on an outcome. Given this fact, the real problem with Nagel and Neef's analysis is that they failed (as did some of the other mathematical modelers we have considered) to make an adequate distinction between individual juror accuracy (first ballots) and jury accuracy (verdicts).

It is appropriate to say that Nagel and Neef's initial analysis assessed weighted errors in the first ballot verdicts produced by juries of varying size who used various decision rules. Their analysis is comparable with the binomial analysis in Figure 1 and Gelfand and Solomon's (1973, 1974, 1975) analysis. What Nagel and Neef's method potentially adds to these other analyses is the emphasis on weighting errors in first ballot verdicts. With respect to these weights, we note that changes in the relative weights attached to false acquittals and convictions will affect the optimal jury size and decision rules. (For example, if false convictions are given a weight of 14 rather than 10, the advantage in weighted errors on first ballot verdicts shifts to the 12-member jury.)

Although Nagel and Neef failed to mention several of the shortcomings in their model, they were aware that the model does a poor job of accounting for the effects of group deliberation on the accuracy (and even the distribution) of jury verdicts. They attempted to overcome this difficulty by arguing that the final verdict distribution is a product of both individual (or *independent*) factors and collective factors. Thus, that 64% of the defendants in the Kalven and Zeisel trials (see Table 1) were convicted by unanimous juries is taken as an index of the collective factors, whereas $.64^{1/12}$ is taken as an index of the independent factors. The collective factors plus the independent factors are presumed to be reflected in the fact that after deliberation 67.7% of the jurors in the 225 Kalven and Zeisel trials voted to convict (when jurors in hung juries are taken into consideration). This 3.7% difference (67.7% − 64%) is, according to Nagel and Neef, attributable to a weighted combination of individual and collective factors:

$$.677 = \left[\frac{W(.64)^{1/n} + .64}{W + 1}\right],$$

where $W = .13$ and $n = 12$. Nagel and Neef's analysis assumed that this relative weighting of independent and collective factors (11% independent and 89% collective) prevails across all decision rules and jury sizes.

We are troubled by this conception of the deliberation process and the calculations to

Table 7
Jury Verdict Errors for Objectively Innocent and Objectively Guilty Defendants

Jury decision rule[a]	% of total no. convictions		% of total no. acquittals		Unweighted sum of errors	Weighted sum of errors[b]
	Correct	Incorrect	Correct	Incorrect		
12/12	60.60	3.19	1.583	30.07	320.6	619.7
8/12	61.71	3.25	1.577	29.95	332.0	624.5
6/6	59.26	3.12	1.647	31.29	344.1	624.9
4/6	60.64	3.19	1.644	31.24	344.3	631.4

[a] The denominator of the decision rule indicates jury size, and the numerator indicates the number of votes needed for a decision.
[b] The weight ratio was 10:1.

which it gives rise. First, it is not clear what this "combined" model really captures. The relative weight of the independent factor is determined solely by the relative imbalance of juror votes in hung juries (13 of the 225 juries in the Kalven and Zeisel trials) averaged over all 225 juries. (Actually, it is not clear that Nagel & Neef's analysis was based on the distribution of votes in the 13 hung cases reported in Table 1; although they implied that they used these 13 cases, their footnotes indicate they may have drawn on a different sample of 48 hung juries.) Irrespective of their data source, it is clear that Nagel and Neef's analysis depends on the level of disagreement in hung juries—if there were no hung juries (or if in all hung juries the proportion of jurors favoring guilt were the same as the proportion of juries who convicted), there would be no independent factor; the combined probability would be identical to the collective probability, and W would equal zero. What Nagel and Neef have done is to take a small bias (toward conviction) in the distribution of votes in a very small number of cases and argue that this bias reflects the individual juror's imperviousness to group influence. Furthermore, by assuming that this independence factor accounts for 11% of the decision making in all jury sizes and decision rules, they have ruled out the possibility that differing jury sizes and decision rules operate directly to affect the extent of group influence on individual jurors (and thereby to affect the distribution of verdict errors). In fact, the evidence that jury size can affect hanging rates (e.g., Kalven & Zeisel, 1966; Padawer-Singer & Barton, Note 1) suggests that the relative impact of independent factors varies with jury size. (Though it is perhaps a less telling defect, Nagel and Neef did not include the possibility of hung juries being produced by their model.)

In addition to the conceptual problems of the combined model, Nagel and Neef's analysis of verdicts is also based on a binomial model in which the probability of a conviction equals the probability that a quorum will be obtained in Y binomial trials. The verdicts of nonquorum juries are once again treated as acquittals, even though the juries may fall only one vote short of a quorum (many such cases end up being added to the error cases for false acquittals).

Although we have chosen not to report the results of Nagel and Neef's verdict analysis (for the reasons enumerated above), we do applaud their efforts and think that with a better formulated verdict model, it would be worthwhile to pursue and test their assumptions about objective guilt and innocence and the effects that different jury sizes and decision rules have on error rates.

One possible approach might make use of the model, developed earlier in this section, that assumes the weight of trial evidence is bimodally distributed. In Figure 2 we used the binomial expansion to generate a distribution of first ballot votes and then applied the decision schemes shown in Table 5 to assess the error rates in final verdicts. If, in common with Nagel and Neef, we are interested in testing inferences about the effects of jury size and decision rule on the weighted probability of false convictions of truly guilty defendants and false acquittals of truly guilty defendants, we can use the same model and, by making

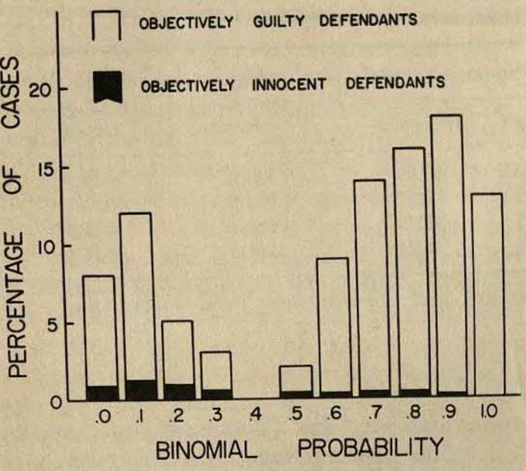

Figure 5. Hypothetical distribution across cases of evidentiary weight for conviction, incorporating the Nagel and Neef (1975) assumption that the distribution of weights is dependent on objective guilt or innocence of the defendant.

reasonable assumptions about the distribution of evidentiary weight against objectively guilty and innocent defendants, assess the probabilities of conviction and acquittal errors for objectively guilty and innocent defendants.

For instance, if we assume (as did Nagel and Neef) that 5% of all defendants are truly innocent and make the further assumption that the evidence against truly guilty and innocent defendants is identically distributed, we can examine the overall distribution of verdicts in Table 6 to assess error rates for different jury sizes and decision rules. As the results in Table 7 demonstrate, with a weight ratio of 10:1, the 12 unanimous juries produced the lowest weighted sum of errors.

A somewhat more complex analysis might incorporate Nagel and Neef's assumption that truly innocent defendants are less likely to be convicted than truly guilty defendants (.4 vs. .7). This assumption can be incorporated into the model by assuming that the evidence against truly innocent defendants is not distributed identically to the evidence against truly guilty defendants, but is skewed in the direction of acquittal. A distribution of evidence such as the one in Figure 5 captures this notion. This distribution of evidence could be used to generate a distribution of verdicts that could then be subjected to an analysis for errors (including an analysis of the weighted sum of errors).

An analysis such as the one advanced here has the advantage of testing the effects of jury size and decision rule variations on error rates (given a variety of assumptions about the distribution of the weight of evidence against objectively guilty and innocent defendants) in the context of a model that straightforwardly relates jury verdicts to initial votes by individual jurors.

Summary Comments on Mathematical Models

As we have noted above, one of the major shortcomings of the mathematical models of jury decision making is the weakness of their assumptions about the relationship between first ballots and final ballots. As one has seen, Saks and Ostrom (1975) did not confront the problem. Walbert (1971) made the simple assumption that verdicts are governed by majority persuasion, with initially evenly divided juries splitting equally for conviction and acquittal, but his model fails to account for reversals of initial majorities and juries that ultimately hang. Gelfand and Solomon (1973, 1974, 1975, 1977) gave little consideration to the relationship of first ballot distributions to final verdicts, for they were able to estimate their parameters simply by examining post hoc, aggregate relationships between first ballots and final verdicts without making assumptions about the intervening processes.

Grofman's (1976) and Nagel and Neef's (1975) models of juror accuracy also suffer from an inability to treat final verdicts, except under the simplest of assumptions about the relationship of first and final ballots.

Davis (1973) and Davis et al. (1975) addressed the problem of modeling social processes and employed a post hoc analysis of first ballots and final verdicts to find the implicit or effective decision rule that best fits the aggregate data. Although good aggregate fits are obtained, no one decision scheme has consistently provided the best fit. Furthermore, when nonaggregated analyses of initial ballots and final verdicts from individual juries have been made, the predictive accuracy of the best fitting models has proven unreliable (Davis, Kerr, Stasser, Meek, & Holt, 1977;

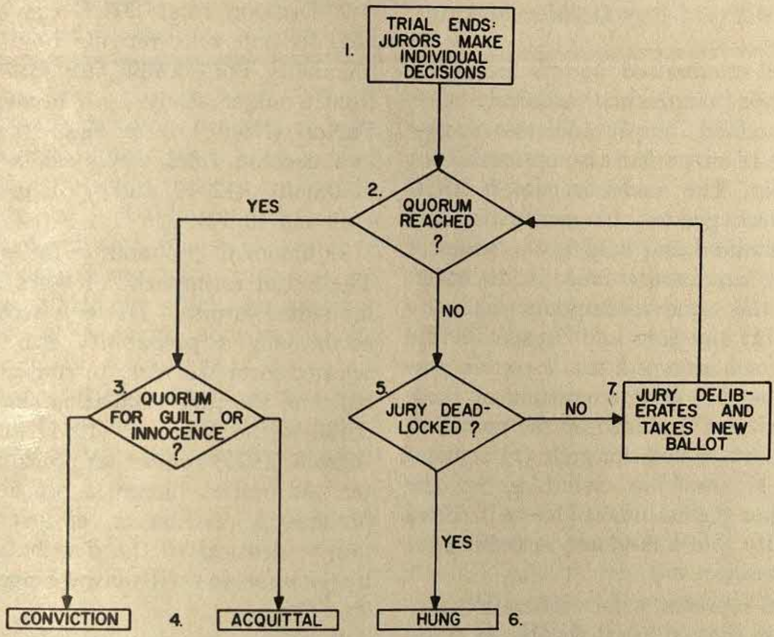

Figure 6. Flowchart summary of our model for jury decision making.

Grofman, 1976; Kerr et al., 1976; Grofman & Hamilton, Note 4).

The binomial model that we have presented avoids some of these problems by introducing more plausible assumptions about the weight of trial evidence and uses a decision scheme approach to determine the distribution of final verdicts. However, its characterization of the deliberation process is still rather barren.

For the present we conclude that the mathematical models built around the binomial theorem are quite adequate for dealing descriptively (and to a lesser extent, predictively) with the relationship between jury size and initial ballot distributions. However, these models are not yet adequate for analysis of jury verdicts. Until we have better knowledge of the actual deliberation process (theoreticians almost universally lament the paucity of relevant data) and this knowledge has been incorporated into the mathematical models, we should be cautious about accepting existing specifications or estimates of the following parameters: the accuracy of juror assessments of guilt and innocence, the prior probabilities of defendants' convictability, the accuracy of jury verdicts, decision rules that maximize juror satisfaction and performance, and the implications of changes in jury size.

We do not mean to imply that these parameters or even the deliberation processes cannot be mathematically modeled, for we are confident that with better and more extensive data, adequate mathematical models can be specified that will provide reliable estimates of the parameters of interest.

In the remainder of this article we summarize the results of our efforts to develop a multiparameter computer model of the deliberation process (the model is presented in detail in Penrod & Hastie, Note 6) that produces output that fits empirical data at a number of critical points (including, but not limited to, reversal and hanging rates). One major advantage of the computer modeling approach is that it allows analysis of the dynamic aspects of the jury decision-making process (i.e., the computer model can easily represent the process by which a jury moves from an initial vote distribution to a final verdict). Furthermore, the simulation method, although compatible with mathematical models, is more flexible in its ability to represent complex hypotheses about juror and jury behavior (e.g., Abelson, 1968).

Computer Model of Jury Decision Making

The model summarized here is somewhat broader in scope than the mathematical models we have examined, but it addresses similar issues and rests on similar assumptions about juror behavior. The model is named DICE after the Greek goddess frequently depicted wearing a blindfold and holding the scales of justice in one hand and a sword in the other. It rests on the same assumptions made by Walbert (1971) and Saks and Ostrom (1975): that at the conclusion of a trial a certain proportion of the jury pool (consisting of those jurors who are not excluded by the voir dire) will be prepared to vote for guilt (p) or innocence ($1 - p$) (or in a civil case, for the plaintiff [p] or the defendant [$1 - p$]). Thus, the probability that a randomly selected juror will vote to convict is p.

The model represents the deliberation process in a form that makes it possible to determine the probability that a randomly selected jury of size n, drawn from a pool in which a specified proportion of the jurors will vote to convict (p), and using any specified decision rule question (q) will produce a conviction, an acquittal, or will hang. Furthermore, the model represents the deliberation process in such a way that it is possible to determine the verdict probabilities for any potential first ballot alignment of votes. In the model decision making is largely characterized by majority persuasion, but in a few cases initial majorities fail to prevail in the deliberations and are either reversed (persuaded by the minority) or do not reach a quorum and hang. Similarly (depending on the case), juries who initially divide evenly sometimes reach verdicts and sometimes hang. Hung juries result when juries fail to attain quorums after extended deliberation. Figure 6 is a simplified representation of the deliberation process embodied in the model, but it does capture the basic structure of DICE.

Parameters

DICE is based on six major parameters:

1. Jury size: Although DICE can operate with any jury size, simulations have concentrated on 6- and 12-member juries.

2. Decision rule: DICE can operate with any decision rule, ranging from majority to unanimity. For example, in modeling the results from a major study by Padawer-Singer and Barton (Note 1), the simulations used the two decision rules employed in that study: unanimity (12/12 and 6/6) and five sixths (5/6 and 10/12).

3. Binomial probability for guilty votes: The initial assignment of votes to simulated individual jurors in DICE is accomplished by establishing a probability that a randomly selected juror will vote to convict on the first ballot of the jury simulation. This parameter parallels the Walbert (1971) and Saks and Ostrom (1975) conviction probability parameter. An initial binomial probability of .47 produces a distribution of first ballot votes nearly identical to the distribution produced by the jurors in the Padawer-Singer and Barton study.

By using a random number generator and the binomial value, all the jurors in a simulation are assigned an initial verdict preference (Step 1).

4. Transition probability function: Several methods of modeling the persuasion/deliberation process have already been noted; Walbert (1971) assumed simple majority persuasion and Davis (1973) offered a wide range of decision schemes. In contrast with these models, which make no attempt to model or explain the deliberation processes that occur between the first and last ballots, DICE follows an alternative approach proposed by Rothschild, Klevorick, and McNeil (Note 7). They have suggested that the deliberation process can be modeled as a continuous-time, birth-and-death Markov process in which the probability that the number of votes for conviction (or acquittal) will increase by one from Time 1 to Time 2 is equal to the proportion of jurors who voted to convict (or acquit) at Time 1. A similar approach has been adopted by Stasser and Davis (1977) and Davis (1978). Experimentation with various transition functions has shown that a function exhibiting a group momentum effect best fits the available empirical data (Penrod & Hastie, Note 6). The transition functions used in DICE are shown in Figure 7. Curve A in Figure 7 shows the probability that a coalition of any size

(from 0–12 in a 12-member jury) will remain intact (i.e., none of the coalition members will change their verdict preference) from one ballot (or time period) to the next. Curve B displays the corresponding probability that individual jurors will not change their verdict preference—the probability values are the *i*th roots of the group probabilities. Briefly, the group transition function (Curve A) captures the following phenomenon: In juries that are roughly equal in size (in which the majority has no more than eight adherents), the majority's persuasive advantage is slight; but as the majority coalition grows in size the probability that it will continue to grow increases exponentially, and single holdouts have a very low probability of not joining the majority (the probability of not changing on any one ballot is .183 for the lone holdout). Readers familiar with early conformity studies (e.g., Asch, 1951; Sherif, 1935) will probably note that these characteristics of the transition function are consistent with conformity research findings. More direct empirical confirmation of the group size effect can be found in research by Godwin and Restle (1974).

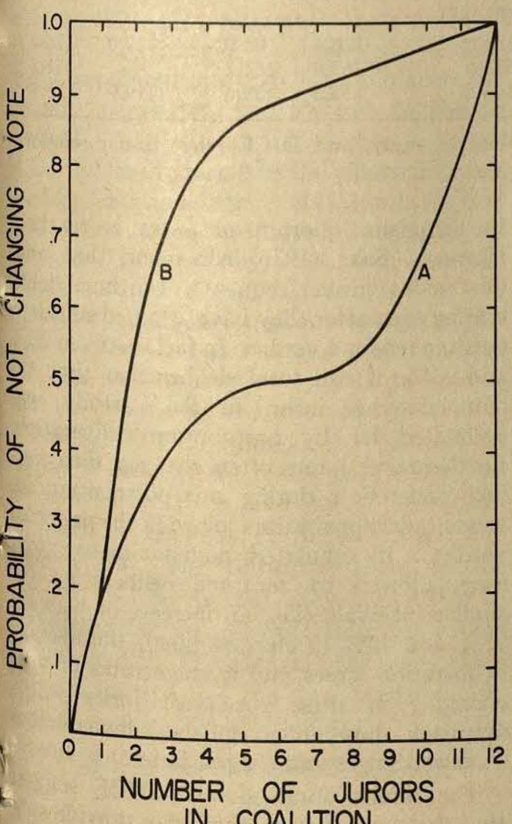

Figure 7. The two functions summarize vote-changing behavior of computer-simulated jurors as a function of coalition size. (Curve A presents the probability that no juror in a coalition of specified size will change votes, and Curve B presents the probability that an individual juror in a coalition will not change votes.)

5. Individual differences: Padawer-Singer and Barton (Note 1) reported that jurors in juries who ultimately reached a verdict were three times more likely to change their verdict preference during deliberation than jurors in juries who hung. This result and other similar results (e.g., the sex difference in change rates in the Kerr et al., 1976, study and the Davis, Kerr, Stasser, Meek, & Holt, 1977, rape case studies) indicate that some jurors are more susceptible to group persuasion than are other jurors. In DICE all jurors are assigned individual *persuasion resistance scores* that reflect these individual differences.

6. Maximum number of ballots: Not all juries reach a verdict; often members feel that they can no longer make progress toward a verdict because individual jurors are entrenched in their preferred verdict. To simulate hung juries DICE sets a limit on the number of ballots that a jury may take.

Evaluative Criteria

The Padawer-Singer and Barton (Note 1) study provides the widest range of data available on jury deliberations and specifically allows us to test DICE's operation with four different criteria: (a) the distribution of jury verdicts (convictions, acquittals, and hung juries); (b) the number of reversals of initial majorities, that is, juries in which a majority of members initially prefer conviction (acquittal) but the jury ultimately renders a verdict for acquittal (conviction); (c) the difference in the rate at which jurors in verdict-reaching juries change their verdict preferences and the rate at which jurors in juries who ultimately hang change their verdict preferences; and (d) the mean and variance in deliberation times for juries who convict, acquit, and hang.

Table 8
Summary of Simulation Results

	12-member jury				6-member jury			
	Unanimous		Nonunanimous		Unanimous		Nonunanimous	
Type of jury	PSB	DICE	PSB	DICE	PSB	DICE	PSB	DICE
				Verdict				
Guilty	8	7.53	9	7.53	8	7.71	9	7.79
Innocent	10	10.70	9	10.70	11	9.99	14	10.13
Hung	5	4.76	5	4.76	4	5.3	0	5.08
				Reversal of majority				
All juries	2	.8	4	.8	1	.3	2	.3
				% of jurors who changed votes				
Verdict juries	34.0	38.4	26.1	23.0	24.6	31.6	28.3	14.9
Hung juries	11.7	15.1	8.4	15.1	8.3	9.6	—	9.6
				Deliberation time				
Verdict juries	169.3	168.0	177.7	134.6	126.4	104.9	119.0	76.9
Hung juries	327.1	322.6	286.7	322.6	253.0	251.2	—	251.2
All juries	203.6	200.9	201.4	173.6	153.3	138.6	119.9	115.4

Note. PSB refers to Padawer-Singer and Barton's (Note 1) study, and DICE refers to our computer model of jury decision making.

Simulation Results

Table 8 summarizes the results of simulations of Padawer-Singer and Barton's 6- and 12-member juries, who used unanimous and five-sixths decision rules, and compares results produced by the actual juries and those obtained from the DICE simulations. Briefly, the distribution of verdicts is quite satisfactory—the poorest fit is in the 6-member, nonunanimous condition, in which none of the actual juries hung. Without exception the simulation juries produced lower reversal rates than the actual juries. This result suggests that the transition probabilities for juries of nearly equal sizes are probably flatter than is reflected in Curve B of Figure 7.

The rate of vote changing in simulated juries is very close to the empirical rate for unanimous juries, but is too low for nonunanimous juries who reach a verdict. Similarly, the average deliberation times for simulated and actual juries are very close for unanimous juries, but the simulated juries produce lower deliberation times in the nonunanimous conditions. At present, DICE renders a verdict as soon as the requisite quorum of votes is reached. However, Saks (1977) has found that nonunanimous juries frequently continue deliberating even after they have attained sufficient votes to render a verdict. In fact, between 20% and 31% of the total deliberation time for nonunanimous juries in Sak's study was accounted for by postquorum deliberation. Furthermore, jurors often changed their verdict preferences during this postquorum interval (additional jurors joined in the preferred verdict). If simulated nonunanimous juries were allowed to continue deliberating for similar intervals (i.e., an increase of between 25% and 40% in elapsed time), the average deliberation times and average rates of vote changing in these simulated juries would approach those found in the nonunanimous Padawer-Singer and Barton juries.

The results obtained with DICE suggest that the simulation method may provide substantial insights into the deliberation process by revealing the relationship between group size and persuasion and by providing a method for assessing the relative impact of individual differences on the persuasion process.

The simulation approach can serve as a useful complement to the mathematical models discussed earlier. DICE provides an alternative method for exploring the implicit decision rules that have been studied by Davis and his colleagues and has the advantage of making explicit and testable assumptions about events that occur between the first ballot and final verdicts. The aspect of the deliberation process that most closely approximates Davis's implicit rule is embodied in the transition function and can be summarized by saying that a majority's persuasiveness increases exponentially as the size of the majority increases. This general rule is likely to prevail across juries and cases, but the model will no doubt require modification when a case produces unusual individual reactions (e.g., DICE has produced excellent fits to the Davis, Kerr, Stasser, Meek, & Holt, 1977, rape data when the differential rates of vote changing by males and females are incorporated into DICE's individual differences parameter (see Penrod & Hastie, Note 6).

DICE also complements the binomial models used by Saks and Ostrom (1975) and Walbert (1971), insofar as the initial distribution of votes in DICE is produced by a binomial function. Furthermore, DICE avoids Walbert's simplistic assumptions about majority persuasion effects and uses a number of criteria to evaluate the assumptions about persuasion that are implicit in the transition function.

Finally, DICE promises to provide an empirical basis for the mathematical analyses of juror error rates and juror satisfaction proposed by Grofman (1976) and Nagel and Neef (1975). The principal defect of the existing analyses is that they are unable to make a satisfactory transition from first ballots to final verdicts. DICE makes the relationship between initial votes and final verdicts quite explicit and therefore provides a basis for extending the existing mathematical models.

Reference Notes

1. Padawer-Singer, A., & Barton, A. H. *Interim report: Experimental study of decision making in the 12- versus 6-man jury under unanimous versus non-unanimous decisions*. New York: Columbia University, Bureau of Applied Social Research, 1975.
2. Penrod, S. *Jury simulation research: Defendant and juror characteristics*. Unpublished manuscript, Harvard University, 1976.
3. Latane, B., & Borden, R. *Theory of social impact*. Research proposal submitted to the National Science Foundation, April 1973.
4. Grofman, B., & Hamilton, L. *Group decision-making in three-member and five-member mock juries under unanimous and non-unanimous verdict requirements*. Unpublished manuscript, State University of New York at Stony Brook, 1976.
5. Nagel, S. S., Lamm, D., & Neef, M. *Decision theory and juror decision-making*. Paper presented at the meeting of the International Society for Political Psychology, New York, September 1978.
6. Penrod, S., & Hastie, R. *Computer simulation of jury decision-making*. Manuscript submitted for publication, 1977.
7. Rothschild, M., Klevorick, A., & McNeil, D. Personal communication, November 1975.

References

Abelson, R. Simulation of social behavior. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, 2nd ed.). Reading, Mass.: Addison-Wesley, 1968.

Asch, S. E. Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men*. Pittsburgh, Pa.: Carnegie Press, 1951.

Badger, W. W. Political individualism, positional preferences, and optimal decision-rules. In R. G. Niemi & H. F. Weisberg (Eds.), *Probability models of collective decision making*. Columbus, Ohio: Charles E. Merrill, 1972.

Boehm, V. R. Mr. Prejudice, Miss Sympathy, and the authoritarian personality: An application of psychological measuring techniques to the problem of juror bias. *Wisconsin Law Review*, 1968, *3*, 734–747.

Cornish, W. R., & Sealy, A. P. Juries and the rules of evidence. *Criminal Law Review*, 1973, April, 208–223.

Curtis, R. B. Decision-rules and collective values in constitutional choice. In R. G. Niemi & H. F. Weisberg (Eds.), *Probability models of collective decision making*. Columbus, Ohio: Charles E. Merrill, 1972.

Davis, J. H. Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, 1973, *80*, 97–125.

Davis, J. H. Group decision and procedural justice. In M. Fishbein (Ed.), *Progress in social psychology*. Hillsdale, N.J.: Erlbaum, 1978.

Davis, J. H., Bray, R. M., & Holt, R. The empirical study of social decision processes in juries. In J. Tapp & F. Levine (Eds.), *Law, justice, and the individual in society: Psychological and legal perspectives*. New York: Holt, Rinehart & Winston, 1977.

Davis, J. H., Kerr, N. L., Atkin, R. S., Holt, R., & Meek, D. The decision processes of 6- and 12-person mock juries assigned unanimous and two-thirds majority rules. *Journal of Personality and Social Psychology*, 1975, *32*, 1–14.

Davis, J. H., Kerr, N. L., Stasser, G., Meek, D., & Holt, R. Victim consequences, sentence severity, and decision processes in mock juries. *Organizational Behavior and Human Performance*, 1977, *18*, 346–365.

Davis, J. H., Kerr, N. L., Sussman, M., & Rissman, A. K. Social decision schemes under risk. *Journal of Personality and Social Psychology*, 1974, *30*, 248–271.

Feinberg, W. E. Teaching the Type I and Type II errors: The judicial process. *American Statistician*, 1971, *25*, 30–32.

Friedman, H. Trial by jury: Criteria for convictions, jury size, and Type I and Type II errors. *American Statistician*, 1972, *26*, 21–23.

Gelfand, A. E., & Solomon, H. A study of Poisson's models for jury verdict in criminal and civil trials. *Journal of the American Statistical Association*, 1973, *68*, 241–278.

Gelfand, A. E., & Solomon, H. Modeling jury verdicts in the American legal system. *Journal of the American Statistical Association*, 1974, *69*, 32–37.

Gelfand, A. E., & Solomon, H. Analyzing the decision-making process of the American jury. *Journal of the American Statistical Association*, 1975, *70*, 305–309.

Gelfand, A. E., & Solomon, H. An argument in favor of 12-member juries. In S. S. Nagel (Ed.), *Modeling the criminal justice system*. Beverly Hills, Calif.: Sage, 1977.

Godwin, W. F., & Restle, F. The road to agreement: Subgroup pressures in small group consensus processes. *Journal of Personality and Social Psychology*, 1974, *30*, 500–509.

Grofman, B. Not necessarily twelve and not necessarily unanimous. In G. Bermant & N. Vidmar (Eds.), *Psychology and the law*. Lexington, Mass.: Heath, 1976.

Grofman, B. Some preliminary models of jury decision making. In C. Tullock (Ed.), *Frontiers of Economics* (Vol. 4). The Hague, Netherlands: Nijhoff, in press.

Kalven, H., & Zeisel, H. *The American Jury*. Boston: Little, Brown, 1966.

Kerr, N. L., et al. Guilt beyond a reasonable doubt: Effect of concept definition and assigned decision rule on the judgments of mock jurors. *Journal of Personality and Social Psychology*, 1976, *34*, 282–294.

Landy, D., & Aronson, E. The influence of the character of the criminal and his victim on the decisions of simulated jurors. *Journal of Experimental Social Psychology*, 1969, *5*, 141–152.

Lempert, R. O. Uncovering nondiscernible differences: Empirical research and the jury-size cases. *Michigan Law Review*, 1975, *73*, 644–707.

Mitchell, H. E., & Byrne, D. The defendant's dilemma: Effect of jurors' attitudes and authoritarianism on judicial decisions. *Journal of Personality and Social Psychology*, 1973, *25*, 123–129.

Nagel, S. S., & Neef, M. Deductive modeling to determine an optimum jury size and fraction required to convict. *Washington University Law Quarterly*, 1975, 933–978.

Poisson, S. D. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Paris: Bachelier, Imprimeur-Librairie, 1837.

Rae, D. Decision-rules and individual values in constitutional choice. *American Political Science Review*, 1969, *63*, 40–56.

Saks, M. J. *Jury verdicts*. Lexington, Mass.: Heath, 1977.

Saks, M. J., & Ostrom, T. M. Jury size and consensus requirements: The laws of probability v. the laws of the land. *Journal of Contemporary Law*, 1975, *1*, 163–173.

Sherif, M. A. A study of some social factors in perception. *Archives of Psychology, N.Y.*, 1935, No. 187, p. 60.

Simon, R. J. *The jury and the defense of insanity*. Boston: Little, Brown, 1967.

Simon, R., & Kaplan, J. K. Latitude and severity of sentencing options, race of the victim and decisions of simulated jurors: Some issues arising from the "Algiers Motel" trial. *Law and Society Review*, 1972, *7*, 87–98.

Simon, R. J., & Mahan, L. Quantifying burdens of proof: A view from the bench, the jury and the classroom. *Law and Society Review*, 1971, *5*, 319–330.

Stasser, G., & Davis, J. H. Opinion change during group discussion. *Personality and Social Psychology Bulletin*, 1977, *3*, 252–256.

Strodtbeck, F. L., James, R. M., & Hawkins, C. Social status in jury deliberations. *American Sociological Review*, 1957, *22*, 713–718.

Taylor, M. Proof of a theorem on majority rule. *Behavioral Science*, 1969, *14*, 228–231.

Tribe, L. H. Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, 1971, *84*, 1329–1393.

Valenti, A. C., & Downing, L. L. Differential effects of jury size on verdicts following deliberation as a function of the apparent guilt of the defendant. *Journal of Personality and Social Psychology*, 1975, *32*, 655–663.

Walbert, T. D. Note: Effect of jury size on probability of conviction—An evaluation of Williams v. Florida. *Case Western Reserve Law Review*, 1971, *22*, 529–555.

Received December 30, 1977

Psychological Bulletin
1979, Vol. 86, No. 3, 493–522

Review and Conceptual Analysis of the Employee Turnover Process

W. H. Mobley, R. W. Griffeth, H. H. Hand, and B. M. Meglino
Center for Management and Organizational Research
University of South Carolina

Research on employee turnover since the Porter and Steers analysis of the literature reveals that age, tenure, overall satisfaction, job content, intentions to remain on the job, and commitment are consistently and negatively related to turnover. Generally, however, less than 20% of the variance in turnover is explained. Lack of a clear conceptual model, failure to consider available job alternatives, insufficient multivariate research, and infrequent longitudinal studies are identified as factors precluding a better understanding of the psychology of the employee turnover process. A conceptual model is presented that suggests a need to distinguish between satisfaction (present oriented) and attraction/expected utility (future oriented) for both the present role and alternative roles, a need to consider nonwork values and nonwork consequences of turnover behavior as well as contractual constraints, and a potential mechanism for integrating aggregate-level research findings into an individual-level model of the turnover process.

Employee withdrawal, in the form of turnover, has sustained the interest of personnel researchers, behavioral scientists, and management practitioners. At the macro level, economists and personnel researchers have demonstrated the relationship between turnover rates and the aggregate level of economic activity, employment levels, and vacancy levels (see, e.g., Armknecht & Early, 1972; Forrest, Cummings, & Johnson, 1977; Price, 1977; Woodward, 1975–1976). At the micro level, behavioral research has established a consistent, although generally weak, correlation between job dissatisfaction and turnover (Brayfield & Crockett, 1955; Locke, 1976; Porter & Steers, 1973; Vroom, 1964; Herzberg, Mausner, Peterson, & Capwell, Note 1). While the economic and job dissatisfaction contributions to turnover are well established, they are conceptually simplistic and empirically deficient bases for understanding the employee turnover process.

Recently, a number of authors (Forrest et al., 1977; Locke, 1976; Mobley, 1977; Porter & Steers, 1973; Price, 1977) have advocated abandoning further replication of bivariate correlates of turnover, particularly job dissatisfaction, in favor of well-developed conceptual models of the turnover process. Such a model is one objective of this article.

Employee turnover is a behavior of interest to many disciplines and is subject to analysis and discussion at many levels of discourse. The approach taken in this article is basically psychological and rests on the belief that turnover is an individual choice behavior. Thus, the individual is the primary unit of analysis. Selecting the individual as the unit of analysis does not mean that turnover research at the unit, organizational, or other aggregate level is not of value and interest. However, to conclude that such studies clarify the individual turnover decision process may be tantamount to what Robinson

The preparation of this article was supported by the Office of Naval Research under Contract N00014-76-C-0938: NR 170-819. Appreciation is expressed to John Logan and John Cathcart for their assistance in the literature search and to the editor and reviewers for their constructive comments.

Requests for reprints should be sent to William H. Mobley, Center for Management and Organizational Research, University of South Carolina, Columbia, South Carolina 29208.

Table 1
Studies of Relations Between Personal Factors and Turnover

Factor	Population	N	Relation to turnover
Age			
Federico, Federico, & Lundquist (1976)	Credit union females	96	Younger age at application associated with lower tenure
Hellriegel & White (1973)	Certified public accountants	349	No differences (significance test not reported)
Mangione (Note 2)	ISR diverse occupational sample	294	Chi-square, $p < .001$; younger age associated with higher turnover
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.22^{**}$; younger age associated with higher turnover
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	$r = -.22^{**}$
Porter, Steers, Mowday, & Boulian (1974)	Psychiatric technicians	60	Stayers significantly older than leavers**
Waters, Roach, & Waters (1976)	Insurance company clerical	105	$r = -.25^{*}$
Tenure			
Mangione (Note 2)	ISR diverse occupational sample	295	Chi-square, $p < .001$; lower tenure associated with higher turnover
Mobley et al. (1978)	Hospital employees	203	$r = -.25^{**}$
Waters et al. (1976)	Insurance company clerical	105	$r = -.30^{**}$
Sex			
Mangione (Note 2)	Institute for Social Research diverse occupational sample	293	*ns*
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.31^{**}$; women had higher turnover
Family responsibilities			
Federico et al. (1976)	Credit union females	96	Higher responsibility associated with higher tenure; factors included marital status, number of children, age of youngest child, and age
Mangione (Note 2)	ISR diverse occupational sample	295	Chi-square, $p < .001$; single people had higher turnover
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.22$; no or few dependents had higher turnover
Waters et al. (1976)	Insurance company clerical	105	*ns* (marital status)
Education			
Federico et al. (1976)	Credit union females	96	Higher education associated with lower tenure
Hellriegel & White (1973)	Certified public accountants	349	No differences (significance test not reported)
Mangione (Note 2)	ISR diverse occupational sample	294	*ns*

(1950) has termed the *ecological fallacy*. For example, the relationship between aggregate unemployment levels and turnover rates, although well established (see, e.g., Armknecht & Early, 1972; Price, 1977; Woodward, 1975–1976) adds little to understanding individual turnover decisions. A linking mechanism is needed that considers the individuals's perception and evaluation of available alternatives relative to the present position.

At the individual level, satisfaction is the most frequently studied psychological variable thought to be related to turnover. However, the satisfaction–turnover relationship, although consistent, usually accounts for less than 16% of the variance in turnover (Locke, 1976; Porter & Steers, 1973). It is apparent that models of the employee turnover process must move beyond satisfaction as the sole explanatory variable.

Recently, the constructs of organizational commitment (Porter, Crampon, & Smith, 1976; Porter, Steers, Mowday, & Boulian, 1974; Steers, 1977), organizational attachment (Koch & Steers, 1978), role attachment (Graen, 1976; Graen & Ginsburgh, 1977), and behavioral intentions (Kraut, 1975; Mobley, 1977; Newman, 1974) have been offered as explanatory concepts in the turnover process. However, the conceptual and empirical identity of these concepts and their interrelationships have not always been clear. An additional objective of the present article is to attempt to clarify and integrate these concepts into a general model of the individual employee turnover process.

A third objective of this article is to update earlier reviews of the literature. The last major review of turnover from the individual perspective was that of Porter and Steers (1973). For somewhat more limited reviews of certain aspects of turnover, see Goodman, Salipante, and Paransky (1973) on the hardcore unemployed and retention and Pettman's (1973) partial review of the March and Simon (1958) model. More recently, Price (1977), a sociologist, has published a significant book that seeks to codify the turnover literature from a variety of disciplines and cultures. The Price work contains a number of references generally not

Table 1 *(continued)*

Factor	Population	*N*	Relation to turnover
Personality Hines (1973)	New Zealand Entrepreneurs Engineers Accountants Middle managers	 80 74 68 93	Entrepreneurs had higher need for achievement and lower turnover rate than other occupational groups; individuals with higher need for achievement had higher turnover in nonentrepreneur groups
Distance migrated Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.12^{**}$; employees who grew up near the factory had higher turnover
Number of previous jobs Marsh & Mannari (1977)	Japanese electrical company employees	1,033	*ns*

Note. ISR = Institute for Social Research.
$^{*}p < .05$. $^{**}p < .01$.

included in the psychological and management turnover literature cited in the United States; however, it does not deal with post-1974 research and is incomplete in its coverage of the psychological and management literature on employee turnover. Forrest et al. (1977) also recently presented a partial review of the turnover literature. However, the latter review, which deals with a broader spectrum of organizational participation behaviors, contains no post-1973 research and has a conceptual rather than an empirical emphasis. The Forrest et al. model is discussed in a later section of this article.

In summary, the major objectives of this article are (a) to update the last major reviews and analyses of the turnover literature, (b) to attempt to clarify the distinctions among various constructs that have recently been suggested as explanatory variables in the turnover process, (c) to develop a conceptual model of the individual-level employee turnover process that is consistent with the research literature, and (d) to suggest areas of further research.

This article focuses on voluntary, that is, self-initiated, turnover rather than on organization-initiated terminations. This distinction, discussed in a subsequent section, is not always made clear in specific research studies. Additionally, this article does not deal with absenteeism. Whether absenteeism is best thought of as having no consistent relationship to turnover (March & Simon, 1958), as a precursor of turnover (Herzberg et al., Note 1), or as an alternative form of withdrawal behavior (Hill & Trist, 1955; Rice & Trist, 1952) is an important research question (Burke & Wilcox, 1972), but is beyond the scope of the present article.

Update of Turnover Analyses and Reviews

The last major reviews of the turnover literature were by Porter and Steers (1973) and Price (1977). This section summarizes the recent research not included in these reviews and offers the conclusions of the authors of this article. A subsequent section summarizes and integrates the results of this and the two previous reviews. Although no taxonomic schema is entirely satisfactory, the research summary is divided into the following sections: (a) individual demographic and personal variables, (b) overall satisfaction, (c) organizational and work environment factors, (d) job content factors, (e) external environment factors, (f) occupational groupings, (g) recently developed constructs, and (h) multivariate studies. Most studies reviewed take a bivariate approach to turnover, but this emphasis is reflective of the current literature rather than of the present authors' belief in the relative merit of this approach.

Individual Demographic and Personal Factors

Included in this category are age, tenure, sex, family responsibilities, education, personality, other personal considerations, and weighted application blanks. Table 1 summarizes recent research on these variables, which were not included in the Porter and Steers or Price reviews.

Age. Recent research, with the exception of Hellriegel and White (1973) who reported no differences, indicates a negative relationship between age and turnover. However, the amount of variance explained is less than 7%. One should note that since age is correlated with many other variables, it alone contributes little to the understanding of turnover behavior. As is noted later, a conceptual model and multivariate studies are required to adequately comprehend the psychology of the turnover process. This observation also applies to each variable discussed below.

Tenure. Three recent studies, cited in Table 1, showed a negative relationship between tenure and turnover. Mangione (Note 2) concluded on the basis of a multivariate study (see Table 12 for a summary of multivariate studies) that length of service is one of the best single predictors of turnover.

Sex. Of the two studies relating an individual's sex to turnover, Marsh and Mannari (1977) observed that female Japanese manufacturing employees had higher turnover than males. Mangione (Note 2) found no relationship.

Family responsibilities. Three of the four studies summarized in Table 1 indicate that family responsibility, including marital status, is associated with decreased turnover.

Education. Of the recent studies dealing with education, one found that female credit union employees with higher education had lower tenure (Federico, Federico, & Lundquist, 1976), whereas Mangione (Note 2) and Hellriegel and White (1973) discovered no differences. Lack of variance in education in studies such as Hellriegel and White's, which used certified public accountants, precludes adequate evaluation of the role of education.

Weighted application blanks. Weighted application blanks use a procedure for weighting information on an employment application form so as to predict some aspect of job performance, including turnover. Schwab and Oliver (1974) have raised serious questions regarding the utility of weighted application blanks for predicting turnover. In four samples, they found that validities shrunk below statistical significance upon cross-validation. However, in two recent studies Cascio (1976) and Lee and Booth (1974) reported significant relationships that were cross-validated. The utility of weighted application blanks for employee selection continues to require situation-specific validation (and regular cross-validation), and alone they offer little contribution to understanding the psychology of the turnover process.

Other personal variables. Table 1 cites other studies that dealt with personality, distance migrated, and number of previous jobs. Because the number of studies is small, no generalizations are possible.

Overall Job Satisfaction and Turnover

Studies involving overall job satisfaction are summarized in Table 2. With one exception (Koch & Steers, 1978), these studies indicate a negative relationship between overall satisfaction and turnover. It is important to note, however, that the amount of variance accounted for is consistently less than 14%. As noted in subsequent sections, when satisfaction is included in multiple regressions with variables such as intentions and com-

Table 2
Studies of Relation of Overall Job Satisfaction to Turnover

Study	Population	N	Relation to turnover
Ilgen & Dugoni (Note 3)	Retail clerks; baggers	117	$r = -.31^{**}$ on Minnesota Satisfaction Questionnaire
Koch & Steers (1978)	Nonmanagement entry-level public agency employees	77	$r = -.37^{**}$ for overall satisfaction *ns* (Job Descriptive Index total)
Mangione (Note 2)	Institute for Social Research diverse occupational sample	295	$r = -.13^{*}$; $r = -.22^{**}$; $r = -.25^{**}$ (three overall satisfaction measures)
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.19^{**}$
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	$r = -.21^{**}$
Newman (1974)	Nursing home employees	108	$r = -.16^{*}$
Waters & Roach (1973)	Female insurance company clerical	80 in follow-up study	$r = -.22^{*}$
		117 in second sample	$r = -.27^{*}$

$^{*} p < .05$. $^{**} p < .01$.

Table 3
Studies of Relations Between Pay–Promotion and Turnover

Factor	Population	N	Relation to turnover
Salary, actual and expected			
Federico, Federico, & Lundquist (1976)	Credit union females	96	Higher salary associated with longer tenure (salary strictly performance based); the greater the difference between expected and actual, the shorter the tenure; the higher the expectations, the lower the tenure
Salary increases			
Hellriegel & White (1973)	Certified public accountants	349	Turnovers reported 20% increases in pay on new job
Satisfaction with pay			
Hellriegel & White (1973)	Certified public accountants	349	Turnovers more negative than nonturnovers on attitudes toward pay policies and comparability of salary (significance levels not reported)
Koch & Steers (1978)	Nonmanagement entry-level public agency employees	77	*ns*
Kraut (1975)	Salesmen	911	*ns*
Mangione (Note 2)	Institute for Social Research diverse occupational sample	295	$r = -.16^{*}$
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	*ns*
Newman (1974)	Nursing home employees	108	*ns*
Waters, Roach, & Waters (1976)	Female insurance company clerical	105	*ns*
Satisfaction with promotion, advancement			
Hellriegel & White (1973)	Certified public accountants	349	Turnovers more negative about opportunities than nonturnovers (significance levels not reported)
Koch & Steers (1978)	Nonmanagement entry-level public agency employees	77	*ns*
Kraut (1975)	Salesmen	911	*ns*
Mobley et al. (1978)	Hospital employees	203	*ns*
Newman (1974)	Nursing home employees	108	*ns*
Waters et al. (1976)	Female insurance company clerical	105	*ns*
Perceived chances of promotion			
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.22^{*}$; poorer perceived chances of promotion had higher turnover

* $p < .01$.

Table 4
Studies of Relations Between Supervision and Turnover

Factor	Population	N	Relation to turnover
Satisfaction with supervision			
Hellriegel & White (1973)	Certified public accountants	349	Turnovers had significantly less favorable attitudes (few significance tests reported)
Ilgen & Dugoni (Note 3)	Retail clerks	117	$r = -.23^*$
Koch & Steers (1978)	Nonmanagement entry-level public agency employees	77	*ns*
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	*ns*
Newman (1974)	Nursing home employees	108	*ns*
Waters, Roach, & Waters (1976)	Insurance company clerical	105	*ns*
Leader acceptance			
Dansereau, Cashman, & Graen (1974)	Managers	354	$r = -.11^*$ (and moderated by expectancy of finding comparable alternative, $r = -.21^*$ for high-expectancy group, $n = 98$)
Graen & Ginsburgh (1977)	University service employees	89	Significant main effect ($p < .024$) with resignation at 24 months; leader acceptance defined in terms of leader's flexibility in changing employee job and chances of leader using his or her power to help employees solve work problems

$^* p < .05$.

mitment, its effect on turnover may become nonsignificant (Marsh & Mannari, 1977; Mobley, Horner, & Hollingsworth, 1978).

Organizational and Work Environment Factors

Pay and promotion. Research dealing with pay and promotion is summarized in Table 3. Federico et al. (1976) found that higher salary was associated with longer tenure, whereas higher salary and the difference between expected and actual salary were associated with shorter tenure. Mangione (Note 2) found a significant negative correlation between pay satisfaction and turnover. Hellriegel and White (1973) discovered that "leavers" had more negative attitudes toward pay than "stayers" and also reported significant increases in pay on their new jobs. Evidence from five other recent studies suggests a lack of relationship between pay satisfaction and turnover.

Also evident in Table 3 is the general lack of relationship between satisfaction with promotion and turnover, although Hellriegel and White did find that leavers had more negative attitudes toward promotion than stayers. Marsh and Mannari (1977) reported a correlation of −.22 between perceived chances of promotion and turnover.

Supervision. Table 4 summarizes recent studies relating satisfaction with supervision to turnover. In four of the studies a nonsignificant relationship between satisfaction with supervision and turnover was found. Hellriegel and White (1973) and Ilgen and Dugoni (Note 3) found a significant negative relationship. The study by Graen and Ginsburgh (1977) is of particular interest. Here, leadership was significantly associated with turnover. The leadership variable, however, was not satisfaction with supervision but specific aspects of the leader–member exchange. Contrasted with the conclusions of recent studies using satisfaction with supervision as the independent variable, the Graen and Ginsburgh results suggest the need for more detailed study of the leader–member exchange rather than reliance on generalized supervision affect measures.

Table 5
Studies of Relations Among Group Cohesiveness, Teamwork, and Satisfaction with Co-Workers

Factor	Population	*N*	Relation to turnover
Co-Workers, Teamwork, Team Effectiveness, and Cohesiveness			
Hellriegel & White (1973)	Certified public accountants	349	"Generally" more negative for turnovers (few significance tests reported)
Ilgen & Dugoni (Note 3)	Retail clerks	117	*ns* (co-workers)
Koch & Steers (1978)	Nonmanagement entry-level public agency employees	77	$r = -.21^*$ (co-workers)
Kraut (1975)	Salesmen	911	*ns* (teamwork)
Mangione (Note 2)	Institute for Social Research diverse occupational sample	295	*ns* (co-workers)
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	*ns* (cohesiveness)
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	*ns* (co-workers)
Newman (1974)	Nursing home employees	108	*ns* (co-workers)
Waters, Roach, & Waters (1976)	Insurance company, clerical	105	*ns* (co-workers)

* $p < .01$.

Table 6
Studies of Relations Between Other Organizational and Work Environment Factors and Turnover

Factor	Population	*N*	Relation to turnover
Role status			
Hellriegel & White (1973)	Certified public accountants	349	Perceptions of prestige and social status in the community lower for turnovers than nonturnovers (significance tests not reported)
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.25^{**}$; lower status had higher turnover
Knowledge of organization's procedures			
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.14^{**}$; lower knowledge had higher turnover
Control processes			
Hellriegel & White (1973)	Certified public accountants	349	Turnovers "generally" more negative than stays, especially re performance evaluation (significance tests not reported)
Role pressures			
Hellriegel & White (1973)	Certified public accountants	349	Not significant on items related to scarcity of time
Climate			
Ilgen & Dugoni (Note 3)	Retail clerks	117	*ns*
Satisfaction with company			
Kraut (1975)	Salesmen	911	*ns*
Satisfaction with hours			
Ilgen & Dugoni (Note 3)	Retail clerks	117	$r = -.33^{**}$
Resource adequacy			
Mangione (Note 2)	Institute for Social Research diverse occupational sample	295	$r = -.13^{*}$
Satisfaction with comfort			
Mangione (Note 2)	Institute for Social Research diverse occupational sample	295	$r = -.20^{**}$

$^{*} p < .05$. $^{**} p < .01$.

Table 7
Studies of Relations Between Job Content Factors and Turnover

Factor	Population	N	Relation to turnover
Satisfaction with work itself			
Koch & Steers (1978)	Nonmanagement entry-level public agency employees	77	$r = -.31^{**}$
Kraut (1975)	Salesmen	911	$r = -.14^{**}$; short term, 18 months $r = -.12^{**}$; long term, $5\frac{1}{2}$ years
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	$r = -.20^{**}$
Newman (1974)	Nursing home employees	108	*ns*
Porter, Steers, Mowday, & Boulian (1974)	Psychiatric technicians	60	Only moderate contribution to discriminate functions with organization commitment the strongest variable
Waters & Roach (1973)	Female insurance company clerical	80 in follow-up	$r = .26^{*}$
		117 in another office	$r = -.40^{*}$
Waters, Roach, & Waters (1976)	Female insurance company clerical	105	$r = -.37^{**}$
Amount of work			
Kraut (1975)	Salesmen	911	*ns*
Job autonomy			
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.09^{*}$; low perceived autonomy had higher turnover
Intrinsic satisfaction			
Mangione (Note 2)	Institute for Social Research diverse occupational sample	295	$r = -.22^{**}$ with challenge
Mirvis & Lawler (1977)	Bank tellers	160	$r = -.20^{*}$
Intrinsic motivation			
Mirvis & Lawler (1977)	Bank tellers	160	$r = -.16^{*}$
Herzberg motivator deprivation			
Karp & Nickson (1973)	Black working poor	50	$r = .37^{**}$ with number of jobs held in last 5 years
Role orientation			
Graen, Orris, & Johnson (1973)	University clerical probationary period	62	25% attrition for high role orientation; 59% attrition for low role orientation at the end of 10 months; role orientation defined as perceived relevance of the job for the worker's career

Table 7 *(continued)*

Factor	Population	N	Relation to turnover
Graen & Ginsburgh (1977)	University service employees	89	Significant main effect ($p < .002$) with resignations
Satisfaction with duties and policies			
Ilgen & Dugoni (Note 3)	Retail clerks	117	*ns*
Self-Perception of task-relevant abilities			
Ekpo-Ufot (1976)	Auto assemblers	123	$r = -.27^{**}$ (sum of 6 measures); higher turnover associated with lower self-perceived abilities
Coping behavior			
Ilgen & Dugoni (Note 3)	Retail clerks	117	*ns* (self-evaluation of coping with critical incidents)
Perceived performance			
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.14$; poorer performers had higher turnover

$^{*}p < .05$. $^{**}p < .01$.

Overall, recent studies offer moderate support for the negative relationship between satisfaction with supervision and turnover. However, the number of studies finding no significant relationship between these variables indicates the need to more closely examine the nature of leadership measures, to conduct more microanalyses of the leader–member exchange, and to assess the contribution of supervision in multivariate designs that consider other salient variables.

Peer group relations. Research on peer relations and turnover is shown in Table 5. In seven of the nine studies, no significant results were reported. Koch and Steers (1978) found a significant correlation between satisfaction with co-workers and turnover, but only 4% of the variance was explained. The studies summarized in Table 5 do not support the generalization of a strong relationship between group relations and turnover. Individual differences in variables such as need for affiliation, the role of other variables, for example, required task interaction and external job alternatives and the method of measuring group relations contribute to the difficulty in explicating these findings.

Other variables. Table 6 describes recent studies dealing with a variety of other organizational and work environment factors. Both Hellriegel and White (1973) and Marsh and Mannari (1977) found a negative relationship between preceived status and turnover. This status may have come, in part, from the work role or from organizational affiliation. Knowledge of organizational procedures and perceptions of control processes were each shown to be negatively related to turnover. In three separate studies, role pressures, climate, and satisfaction with the company were not significantly related to turnover. Ilgen and Dugoni (Note 3) found a significant negative correlation between satisfaction with hours of work and turnover among retail clerks. Mangione (Note 2) discovered a significant negative correlation between resource adequacy and turnover.

Job Content Factors

Recently job content has become one of the more active areas of industrial–organiza-

tional research. A number of studies in this area have employed turnover as a criterion. Inspection of the recent studies shown in Table 7 indicates that job content factors are significantly related to turnover. Satisfaction with work itself exhibits a uniform negative correlation with turnover, although the amount of variance explained is consistently less than 16%.

Additional studies indicate that the perceived intrinsic value of work, intrinsic motivation, and intrinsic satisfaction are all significantly and negatively related to turnover. Graen, Orris, and Johnson (1973) and Graen and Ginsburgh (1977) demonstrated that role orientation, defined as the perceived relevance of the job for the worker's career, was significantly related to turnover.

In a particularly interesting study, because it took a somewhat different approach, Ekpo-Ufot (1976) found that self-perception of task-relevant abilities was significantly and negatively associated with auto assembler turnover.

External Environment

The probable role of the availability of alternative jobs in employee turnover has long been recognized; see, for example, March and Simon (1958). Economists and sociologists have documented the aggregate-level relationship between economic indicators such as employment levels or job vacancy rates and turnover rates (Price, 1977). However, research on individual-level turnover has infrequently assessed perceived alternatives (Forrest et al., 1977; Locke, 1976). Conceptually, the perception and evaluation of alternatives seems to be a crucial variable in the individual turnover process. Empirically, assessment of the relationship between turnover and personal, organizational, job content, or other variables is inexorably bound to consideration of the perception and evaluation of alternatives.

Table 8 summarizes the limited amount of recent research dealing with alternatives. Woodward (1975–1976) reaffirmed the aggregate-level negative relationship between unemployment and turnover and the positive

Table 8
Studies of the Relation Between External Alternatives and Turnover

Factor	Population	*N*	Relation to turnover
Economic conditions Woodward (1975–1976)	British chemical company	not reported	Local labor market: unemployment, $R^2 = .34$; unfilled vacancies, $R^2 = .66$; unemployment, vacancy, and a seasonal dummy, $R^2 = .83$
Expectancy of finding alternative Dansereau, Cashman, & Graen (1974)	Office workers Managers	222 354	In both samples, the perceived expectancy of finding a comparable job moderated the correlation between attitude and turnover (office workers, $r = -.18^{**}$; managers, $r = -.11^{*}$)
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	Expectancy of finding an acceptable alternative significantly correlated with intention to quit ($r = .15^{*}$); intention to search correlated significantly with turnover ($r = .29^{**}$)

$^{*}p < .05$. $^{**}p < .01$.

relationship between unfilled vacancies and turnover rates. At the individual level, Dansereau, Cashman, and Graen (1974) found that the expectancy of finding an alternative job moderated the correlations between attitude and turnover. Mobley et al., (1978) found that expectancy of finding an acceptable alternative position was significantly and positively related to intention to quit but not to actual quitting, although intention to quit was significantly and positively related to turnover. It is evident that much additional research is required to explicate the role of perception and evaluation of alternatives in the individual turnover process.

Occupational Groupings

Price (1977) reviewed research on occupational characteristics and found moderate support for the proposition that unskilled blue-collar workers have higher turnover than white-collar workers. He found only weak support for the propositions that nonmanagers have higher turnover than managers, that nongovernment employees have higher turnover than government employees, and that higher professionalism is associated with higher turnover.

Since most individual-level turnover studies are carried out within occupational groupings, the present review adds little to Price's analysis. It is apparent, however, that any complete model of individual turnover behavior should be able to account for differences in turnover among occupational groupings. The perception and evaluation of alternatives is one obvious link between the two levels of analysis. A second is exemplified in the research of Herman and Hulin (1972) and Herman, Dunham, and Hulin (1975), which demonstrated that organizational variables such as position level may be better predictors of behavior than demographic or personality variables. The frame of reference provided by position level may influence values, perceptions, and expectations, thus linking organizational variables with individual behavior.

Recently Explored Variables and Processes

Since the Porter and Steers (1973) review, interest has developed in a variety of additional variables, constructs, and processes, including behavioral intentions, organizational commitment, realistic expectations, and the centrality of work values.

Behavioral intentions. The Fishbein (1967) and Fishbein and Ajzen (1975) model of the relationships among beliefs, attitudes, intentions, and behaviors emphasizes the role of intentions in understanding the link between attitudes and behavior. The Locke (1968) model of task motivation also conceives of intention as an immediate precursor of behavior. Drawing on these and other related theoretical models, a number of recent studies have assessed the role of intentions in predicting and understanding turnover. Table 9 summarizes these studies.

It is evident from these studies that behavioral intentions to stay or leave are consistently related to turnover behavior. It is also evident that this relationship generally accounts for more variance in turnover than does the satisfaction–turnover relationship. Conceptually this appears appropriate, since satisfaction is an affective or emotional response, whereas intentions are statements regarding the specific behavior of interest, in this case, turnover. It is possible, as Mobley et al. (1978) suggested, that intentions also capture the individual's perception and evaluation of alternatives.

Although the relationship between intentions and turnover appears to be consistent and generally stronger than the satisfaction–turnover relationship, it accounts for less than 24% of the variance in turnover. Among the possible reasons for this are that intentions do not account for impulsive behavior, that they do not adequately capture the perception and evaluation of alternatives, and that along with personal, organization, and external conditions, they may change between original measurement and the observation of actual behavior. The more specific the behavioral intention statement and the less time its measurement and the behavior, the stronger the relationship should be. However, as Graen and Ginsburgh (1977) noted,

Table 9
Studies of Relations Between Intentions and Turnover

Factor	Population	N	Relation to turnover
Behavioral intentions			
Kraut (1975)	Salesmen	911	Lower intent to remain significantly associated with turnover after 18 months, $p < .01$; between 18 and 60 months, $p < .05$ ($n = 791$; $r = -.17$* for voluntary turnover in both short- and long-term cases
Mangione (Note 2)	Institute for Social Research diverse occupational sample	242	Intentions significantly related to turnover, $p < .001$ (over 2 years)
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$\tau\text{-}b = -.13$* (4 years)
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	$r = .49$* for intention to quit (47 weeks); $r = .29$* for intention to search; $r = .19$* for thinking of quitting
Newman (1974)	Nursing home employees	108	$r = .39$* (over 2 months)
Waters, Roach, & Waters (1976)	Insurance company clerical	105	$r = -.42$* (over 2 years with intent to remain) $r = .36$* (with biographical data partialed out)
Attitude toward act of quitting			
Newman (1974)	Nursing home employees	108	$r = .30$*
Normative beliefs re quitting			
Newman (1974)	Nursing home employees	108	$r = .32$*

* $p < .01$.

Table 10
Studies of Relations Among Organizational Commitment, Involvement, Job Attachment, and Turnover

Factor	Population	N	Relation to turnover
Organizational commitment			
Marsh & Mannari (1977)	Japanese electrical company employees	1,033	$r = -.09^*$, but no significant contribution in multiple regression with 15 other variables
Porter, Steers, Mowday, & Boulian (1974)	Psychiatric technicians	60	Commitment had largest standardized weight in discriminant function, which included Job Descriptive Index
Porter, Crampon, & Smith (1976)	Management trainees	32	$r = .41^*$ for Day 1 measure; $r = .43^*$ for measure in last 2 months with turnover data over 15 months; decline in commitment prior to actually leaving
Steers (1977)	Hospital employees	382	$r = -.17^{**}$ (1 year); commitment and intention to remain, $r = .31^{**}$; desire to remain, $r = .44^{**}$
Organizational involvement			
Mirvis & Lawler (1977)	Bank tellers	160	$r = -.29^{**}$ (3 months)
Job attachment			
Koch & Steers (1978)	Nonmanagerial entry-level public agency employees	77	$r = -.38^{**}$; $r = -.40^{**}$ with satisfaction partialed out

$^* p < .05.$ $^{**} p \leq .01.$

the more specific the intention measure and the closer the person is to actually quitting, the more trivial the prediction. Additionally, without analyses of the precursors of intentions, little knowledge of the psychology of turnover behavior is generated.

Also included in Table 9 are two variables related to intentions in the Fishbein (1967) model: attitude toward the act of quitting and normative beliefs regarding quitting. The Newman (1974) study is one of the few that tests the Fishbein model with a turnover criterion. Both variables, as well as intentions, were significantly related to turnover.

It is evident that intentions are a significant variable in the turnover process. However, additional research is required on the antecedent and covariates of intentions, the manner in which intentions change over time, and the reasons for the lack of a stronger relationship between intentions and turnover.

Organizational commitment, involvement, and job attachment. A number of researchers have recently focused on the antecedents and consequences of organizational commitment (see, e.g., Porter et al. 1974; Steers, 1977). Porter et al. (1974, p. 604) defined organizational commitment as a more global evaluative linkage between the employee and the organization, which includes job satisfaction among its components. More specifically, organizational commitment was defined as the strength of an individual's identification with and involvement in a particular organization and is characterized by (a) a strong belief in and acceptance of an organization's goals and values, (b) a willingness to exert considerable effort on behalf of the organization, and (c) a definite desire to maintain organizational membership. Porter et al. stated that intention to remain is a component of commitment.

More recently, Koch and Steers (1978) suggested that job attachment may be a primary precursor of turnover. They defined job attachment as an attitudinal response to one's job characterized by (a) a congruence between one's real and ideal jobs, (b) an identification with one's chosen occupation, and (c) a reluctance to seek alternative employment. Koch and Steers further noted that job attachment is clearly related to organizational commitment, although it focuses more specifically on one's occupation or job than on the organization as a whole, that job attachment should be more closely related to turnover than is satisfaction because of its conative or intentional component, and that it should be influenced relatively more by individual than by job characteristics.

Table 10 summarizes recent studies of the relation between these variables and turnover. Porter et al. (1974), Porter et al. (1976), and Steers (1977) all found that commitment was more significantly and negatively related to turnover than was satisfaction. Marsh and Mannari (1977) discovered a significant but weak negative correlation between commitment and turnover among Japanese employees, whereas Mirvis and Lawler (1977) observed that organizational involvement, one component of commitment, was significantly and negatively related to turnover.

Koch and Steers (1978) found job attachment to be significantly and negatively related to turnover. It should be noted that role orientation in the previously reviewed Graen et al. (1973) and Graen and Ginsburgh (1977) studies (perceived relevance of the job for the worker's career) is related to at least one aspect of Koch and Steers's definition of job attachment. The Graen studies found role orientation to be significantly and negatively related to turnover.

The developing body of research on commitment and attachment suggests that these concepts are significantly and negatively related to turnover and more strongly related to turnover than to satisfaction. However, both commitment and attachment, as defined in the research cited above, are such complex constructs as to make generalizations rather tenuous. For example, is it the inclusion of intentions in the operational definitions of commitment and attachment that accounts for their improved prediction of turnover? Is it not possible for congruence between individual and organizational goals and values to vary independently of the other two components of commitment? Perhaps a more microanalytic treatment of these constructs

Table 11
Studies of Relations Between Met Expectations and Turnover

Study	Population	N	Relation to turnover
Dunnette, Arvey, & Banas (1973)	College graduates	1,020	Leavers had larger differences than stayers on a Vroom-type (1964) motivation index between expectations and job experiences on last job (significance levels not reported)
Farr, O'Leary, & Bartlett (1973)	Female sewing machine operators	160	White applicants administered a work sample test had lower voluntary turnover than group given traditional tests, $p < .05$; no differences for minority applicants (no direct measure of expectations)
Ilgen & Seely (1974)	New West Point cadets	468	Those receiving booklet of realistic information had lower turnover than control group (no direct measure of expectations and treatment after acceptance decision)
Ilgen & Dugoni (Note 3)	Retail clerks	117	Met expectations was inconsistently related to turnover, although summary met expectations variable was significantly and negatively related to turnover ($r = -.22^{**}$)
Wanous (1973)	New telephone company operators	80	Realistic compared with traditional job preview had lower expectations on Job Descriptive Index Work* and Supervision,** lower thoughts of quitting,** and lower measures on relevant Minnesota Satisfaction Questionnaire items and job survival ($p < .20$)

* $p < .05$. ** $p < .01$.

would prove useful. A model that incorporates some components of commitment and attachment is discussed in a subsequent section.

Met expectations. Porter and Steers (1973) suggested that met expectations provide a conceptual framework for the diverse turnover literature. They viewed this concept as the discrepancy between what a person encounters on the job in the way of positive and negative experiences and what was expected. They predicted that "when an individual's expectations—whatever they are—are not substantially met, his propensity to withdraw would increace" (Porter & Steers, 1973, p. 152).

Since the Porter and Steers review, there have been several studies relevant to the met expectations hypothesis. A subset of these studies dealt with expectations at the time of original organizational entry. Table 11 summarizes studies since the Porter and Steers review that are most relevant to met expectations.

The Dunnette, Arvey, and Banas (1973) study found that leavers exhibited a greater discrepancy between original expectations and actual experiences than did stayers. However, significance levels were not reported and leavers' perceptions of their last job were retrospective, suggesting the possibility of postdecision distortion.

Farr, O'Leary, and Bartlett (1973) and Ilgen and Seely (1974) found some evidence that individuals given realistic information about the job (via a work sample and a booklet) exhibited lower turnover. However, in neither study were expectations or subsequent experiences directly assessed.

Ilgen and Dugoni (Note 3) sought to assess directly the met expectation hypothesis, but found that met expectations were inconsistently related to satisfaction or turnover. Wanous (1973) discovered that a realistic job preview, compared with a more traditional orientation, lowered both expectations and thoughts of quitting, but did not significantly influence turnover.

Direct support of the met expectations hypothesis is rather weak. In an insightful discussion of the conceptual and empirical support for met expectations (particularly as related to realistic job previews), Ilgen and Dugoni concluded that to expect realistic job previews to influence satisfaction, and subsequently turnover, through the mechanism of met expectations is naive. Specifically, the hypothesis is theorized to inadequately reflect individual differences in values. However, it should be noted that Porter and Steers (1973) appeared to account for individual differences through "desired expectations." Ilgen and Dugoni also noted that accurate expectations cannot compensate for deficiencies in the immediate job environment. Additionally, the met expectations hypothesis appears to give insufficient attention to the socialization and assimilation processes.

Although expectations may play an important role in attachment, satisfaction, and turnover, a more complex conceptualization than the met expectations hypothesis appears necessary. One such conceptual model is proposed in a subsequent section of this article.

Multivariate Studies

The preceding sections of this review have repeatedly suggested that multivariate studies are necessary in turnover research. Such studies are necessary in order to interpret the relative efficacy of numerous variables and constructs thought to be related to turnover, to resolve apparently contradictory bivariate studies, to attempt to account for a greater proportion of the variance in turnover, and to move toward a more complete understanding of the turnover process.

Table 12 summarizes recent multivariate studies that have used turnover as the criterion. Graen and Ginsburgh (1977) found that role orientation, leader acceptance, and their interaction accounted for 23% of the variance in university service employee turnover. That the interaction accounted for 6% of the variance suggests that noncompensatory models of turnover may be required.

Mobley et al. (1978) tested a simplified version of a model of possible intermediate linkages between job satisfaction and turnover (Mobley, 1977). Although a number of

Table 12
Summary of Recent Multivariate Studies of Turnover

Study	Population	N	Independent variables and relation to turnover
Graen & Ginsburgh (1977)	University service employees	89	Significant main effects for role orientation (10% variance explained) and leader acceptance (7% variance); significant interaction (6% variance), with resignations at 24 months
Mobley, Horner, & Hollingsworth (1978)	Hospital employees	203	Intention to quit was the only significant regression coefficient ($r^2 = 24\%$) in equation that included intent to search, thinking of quitting, probability of finding acceptable alternative, age, tenure, and satisfaction (47 weeks)
Mangione (Note 2)	Institute for Social Research diverse occupational sample	295	Using 15 demographic, satisfaction, and occupational variables, $R = .63$ (shrunken $r^2 = 22\%$); rank ordered by betas, satisfaction with comfort, satisfaction with co-workers, industry, age, tenure, occupation, satisfaction with financial rewards, occupational prestige, satisfaction with challenge, education, marital status, resource adequacy, race, collar, and sex
Marsh & Mannari (1977)	Japanese electrical company employees	784	$R = .34$ (16 variables); sex, $-.27^{**}$; organizational status, $-.08^{*}$; performance, $.07^{*}$; number previous jobs, $.06^{*}$ (betas); nonsignificant variables: promotion chances, cohesiveness, participation, distance migrated, knowledge of procedures, values, size of community of origin, lifetime commitment, residence, autonomy, and satisfaction
Newman (1974)	Nursing home employees	108	$R = .48^{**}$ for JDI scales, faces scale, attitude toward quitting, normative beliefs regarding quitting, and intentions to quit (not cross-validated; individual betas not reported)
Porter, Steers, Mowday, & Boulian (1974)	Psychiatric technicians	60	Organizational commitment and JDI with age partialed out; significant discriminant function at 2 times closest to actual termination (20.7%) and 21% variance related to termination decision
Waters, Roach, & Waters (1976)	Insurance company clerical	105	$R = 50$, with intentions, JDI work, and tenure the only significant variables; age, other JDI scales, and marital status not significant; intentions accounted for 18% of variance; JDI work and tenure added 7%

Note. JDI = Job Descriptive Index. $^{*}p < .05$. $^{**}p < .01$.

demographic, satisfaction, and perceived alternative measures exhibited significant bivariate relations with turnover, multiple regression analysis revealed intention to quit as the only significant coefficient ($r^2 = 24\%$).

Mangione (Note 2) used 15 demographic, satisfaction, and occupational variables to predict turnover. He found the strongest regression coefficients to be satisfaction with comfort, satisfaction with co-workers, rewards, industry type, age, tenure, occupation, and satisfaction with financial rewards ($r^2 = 40\%$; adjusted $r^2 = 22\%$). Of particular interest was the fact that three different classes of variables (satisfaction, demographic, and occupational) were represented among the strongest regression coefficients. Satisfaction did not subsume the unique variance in the demographic and occupational variables. Unfortunately, this study did not include perception and evaluation of alternatives.

The Marsh and Mannari (1977) study is of particular interest because it dealt with a Japanese sample. These authors found only four variables with significant coefficients (sex, organizational stature, performance, and number of previous jobs). Commitment and satisfaction were among those that did not exhibit significant regression coefficients. This study serves to emphasize the necessity of evaluating models that can generalize beyond the United States and Western industrialized nations.

Newman (1974) conducted one of the few direct tests of the Fishbein (1967) model. Although individual regression coefficients were not reported, he discovered that 23% of the variance in turnover was accounted for by satisfaction, attitude toward quitting, normative beliefs regarding quitting, and actual intentions to quit. The multivariate study by Waters, Roach, and Waters (1976) found a coefficient of 25%. Intentions accounted for 18%, whereas the Job Descriptive Index (JDI) Work scale (Smith, Kendall, & Hulin, 1969) and tenure added the additional 7%.

Porter et al. (1974) observed that organizational commitment and the JDI accounted for 21% of the variance in turnover at two different points in time with age partialed out. Commitment made the stronger contribution.

Several generalizations are possible from these studies. First, each of the studies accounted for more variance in turnover than did satisfaction or any other of the single variables. Thus, satisfaction does not appear to be an adequate composite of other precursors and correlates of turnover. Also, it is evident that intentions, whether measured directly or included in commitment, enhance the prediction of turnover. With the exception of the Mobley et al. (1978) study, other variables when combined with intentions enhanced the prediction. It should be noted that few of the multivariate studies included either perception and evaluation of alternatives or cross-validation. These omissions continue to be major shortcomings of the research.

Summary

Employee turnover remains a frequently researched phenomenon. This is evident from the number of studies since the Porter and Steers (1973) review. Many of these studies have dealt with only a small subset of the variables potentially relevant to turnover, and many are not based on a clear conceptual model. This precludes making strong summary generalizations of the research studies.

Table 13 presents a summary of the Porter and Steers (1973) review, the Price (1977) review, and the present authors' conclusions based on the recent research. In the placement of categories in Table 13 an attempt has been made at maintaining the integrity of the various authors' classification schema and at calling attention to possible overlap in classification groupings. In interpreting the table, *negative* refers to a negative relationship, that is, the higher the variable the lower the turnover, while *positive* refers to a positive relationship. In the case of normal variables, the nature of the relationship is specified.

The qualifiers *consistent*, *moderate*, *weak*, or *inconclusive* are used in Table 13. These qualifiers refer to the consistency with which

a significant relationship was found and to the relative number of studies reporting such a relationship. These qualifiers do not refer to the strength of a relationship in terms of the size of a correlation or variance explained.

The present review, in agreement with the earlier reviews of Porter and Steers and Price, found age, tenure, overall job satisfaction, and reaction to job content to be consistently and negatively associated with turnover. Among the more recently studied variables, intentions and commitment–attachment were found to consistently relate to turnover. Because of the relatively few multivariate studies, an ordering of these variables in terms of relative contribution to turnover is tenuous. However, it appears that intentions and commitment–attachment (which includes intentions) made a stronger contribution to turnover behavior than did satisfaction and demographic variables. Further research is needed for an adequate mapping of the antecedents of intentions.

Moderately consistent support for the negative relationship between supervisory style and turnover was evident, a somewhat more qualified conclusion than that reached by Porter and Steers (1973). Recent research reveals an inconclusive pattern of results with respect to pay, promotion, and peer group relations. These results stand in contrast with the consistent negative generalization of Porter and Steers's review. Differences in the availability of alternatives, the lack of multivariate studies, the lack of multiple measures of perception or affect, and the lack of a clear conceptual model make interpretation of these differences difficult.

The compelling conceptual argument that alternatives are an important variable in the turnover process continues to be supported in aggregate-level studies, but has weak support at the individual level because it has been infrequently studied.

Direct support for the met expectations hypothesis is weak. Although realistic job previews have been shown to be a possible aid in reducing turnover, the psychology of this effect is not well understood.

Finally, the limited number of multivariate studies indicates that greater variance in turnover can be explained by using multiple variables, that a great deal of variance is still unexplained, that inclusion of intentions significantly enhances the prediction of turnover, and that satisfaction is an inadequate summary variable for capturing the effects of other demographic, organizational, occupational, or external variables.

Methodological and Conceptual Comments

Predictive Designs

It is encouraging to note that an overwhelming majority of recent research designs have been predictive rather than retrospective. However, few studies have used repeated measures of the perceptual or affective variables, the Porter et al. (1976) and Graen and Ginsburgh (1977) studies being substantive exceptions. To the extent that turnover is a dynamic process, longitudinal designs with repeated measures should be of high utility.

Linear Models

Most research has been based on the assumptions of a linear and compensatory model. Fleishman and Harris (1962) earlier demonstrated the possibility of a nonlinear relationship between supervisory style and turnover. Mangione (Note 2) was one of the few authors to examine possible nonlinear relationships. Additionally, the Graen and Ginsburgh (1977) finding of a significant interaction between role orientation and leader acceptance calls attention to the need for further exploration of interaction terms.

Criterion

A troublesome issue in turnover research concerns the definition of turnover rates and types of turnover at both the aggregate and individual levels (see Price, 1975–1976, 1977, for an evaluation of a number of aggregate measures of turnover). At the individual level, one of the more troublesome issues is the distinction between voluntary and involuntary turnover. The bulk of the individual-

Table 13
Summary of Three Reviews of the Turnover Literature

Variable	Porter & Steers (1973)	Price (1977)	Present review
Personal characteristic			
Age	Consistent negative	Consistent negative	Consistent negative
Tenure	Consistent negative	Consistent negative	Consistent negative
Similarity of job with vocational interests	Weak negative		
Personality	Weak negative for extreme traits		
Family size and responsibilities	Generally positive for women; inconclusive for males		
Sex		Inconclusive	Inconclusive
Education		Weak positive	Inconclusive
Weighted application blank			Moderate positive
Overall job satisfaction	Consistent negative	Consistent negative	Consistent negative
Organizational and job characteristic			
Pay	Consistent negative	Consistent negative	Inconclusive
Promotion	Consistent negative	Weak negative	Inconclusive
Size of organization	Inconclusive	Inconclusive	
Size of work unit	Consistent positive for blue collar; inconclusive for white collar	Inconclusive	
Peer group interaction	Moderate negative		Inconclusive
Integration		Consistent negative	
Supervision style	Consistent negative		Moderate negative
Instrumental communication		Consistent negative	
Formal communication		Consistent negative	
Role clarity	Consistent negative		
Job autonomy and responsibility	Consistent negative		
Centralization		Consistent negative	
Task repetitiveness	Moderate positive	Weak positive	
Overall reaction to job content	Consistent negative		Consistent negative
Occupational grouping			
Blue collar: skilled vs. unskilled		Moderate (unskilled higher)	
Blue collar vs. white collar		Moderate (blue collar higher)	
Nonmanagers vs. managers		Weak (nonmanagers higher)	
Nongovernment vs. government		Weak (nongovernment higher)	
Professionalism		Weak positive (professionalism higher)	

Table 13 (*continued*)

Variable	Porter & Steers (1973)	Price (1977)	Present review
External environment			
Level of employment/opportunity		Consistent positive	Consistent positive
Perceived alternatives			Weak positive
Recently studied variable			
Intentions to quit			Consistent positive
Commitment/attachment			Consistent negative
Met expectations			Weak negative

level turnover research focuses on voluntary turnover.

Precise definitions of voluntary turnover are infrequently given, and what is included as voluntary may differ across studies. For example, Marsh and Mannari (1977) incorporated pregnancy under voluntary turnover, whereas Mirvis and Lawler (1977) and Waters et al. (1976) excluded pregnancy. The definition of voluntary may well have contributed to Marsh and Mannari's finding of a significant difference in turnover as a function of sex. Much more subtle effects may be associated with results as a function of the definition of voluntary.

The categorization of reasons for turnover is frequently taken from company records. Many personnel practitioners readily admit that a variety of factors influence the administratively recorded reason for attrition. Lefkowitz and Katz (1969) reported significant differences in administrative and self-reported reasons for termination. In agreement with Forrest et al. (1977), further efforts to clarify the implications of different operational measurements of turnovers are appropriate. A multiple measure approach to identifying reasons for turnover would be useful.

Finally, little research has addressed the relationship between voluntary and involuntary turnovers. To assume that these are completely independent phenomena, especially in the case of discipline-related terminations, appears simplistic.

Although turnover is frequently thought of as a "clean" objective criterion, the issues raised above suggest the need for greater attention to the criterion problem in turnover research.

Measures of Satisfaction

In recent years, the JDI (Smith et al., 1969) has become the predominant measure of satisfaction with various facets of the job setting. The majority of satisfaction–turnover studies reviewed in this article used the JDI. The careful development of the JDI is well documented, and there is a clear advantage to using a common satisfaction measure across a variety of studies. However,

overreliance on any single measure raises the possibility that method variance has contaminated supposedly generalizable relationships. As Gillet and Schwab (1975) suggested, it seems prudent to use multiple measures of the same construct wherever possible.

Time

The role of time in turnover research is evident in a number of ways. As noted earlier, there is a consistent negative relationship between tenure (length of time on the job) and turnover. Some studies have ignored tenure, some have partialed out its effect, and others have included it in a multivariate design. Understanding the psychology of the tenure effect is probably best facilitated by the latter treatment.

The time variable is also part of the criterion problem to the extent that different studies measure turnover over different lengths of time. Marsh and Mannari (1977) collected their turnover data over a 4-year period, whereas other studies have looked at turnover over a matter of weeks, for example, Newman (1974). The effect of differing lengths of time between measurement of independent variables and the turnover behavior is infrequently studied. Porter et al. (1976) and Waters et al. (1976) are exceptions. This appears to be a topic in need of additional research.

Finally, the temporal dimension may be relevant to the extent that different variables or combinations of variables exert a differential influence on turnover as a function of stages in the organizational socialization process. Graen and Ginsburgh (1977) discussed this possibility.

Primacy of Work

Turnover is generally conceptualized in terms of demographic, organizational, and individual affective factors and on infrequent occasions in terms of perceived alternatives. While such conceptualizations may reflect individual values relative to the work setting, they do not reflect the importance of work-related values relative to other life values and interests. The work of Dubin and his associates (see, e.g., Dubin, Champoux, & Porter, 1975) has demonstrated that differences in central life interests are related to differences in evaluations of the work environment and in organizational commitment. Marsh and Mannari (1977) found a significant negative relationship between primacy of work values and turnover. It appears that future turnover research should deal not only with the work environment and external alternatives but also with the centrality of work relative to other life values and interests.

Conceptual Model of Employee Turnover

Drawing in part on the review and analysis presented earlier, this section develops a conceptual model of the employee turnover process. A simplified schematic representation of this model is presented in Figure 1. Among the characteristics of this model are the following:

1. It is a model of individual-level turnover behavior. Individual differences in perceptions, expectations, and values are explicitly recognized. Further, individual differences in personal and occupational variables are included.

2. Perception and evaluation of alternative jobs is given explicit treatment.

3. The probable roles of centrality of work values and interests relative to other values and interests, beliefs regarding nonwork consequences of quitting or staying, and contractual constraints are specifically recognized.

4. The possible joint contribution to turnover of job satisfaction (present affect), job attraction (expected future affect), and attraction of attainable alternatives is proposed.

5. Intention to quit is considered to be the immediate precursor of turnover, with impulsive behavior and the time between measurement of intentions and behavior attenuating this relationship.

The rationale for the model is described starting with turnover behavior and working back through its antecedents.

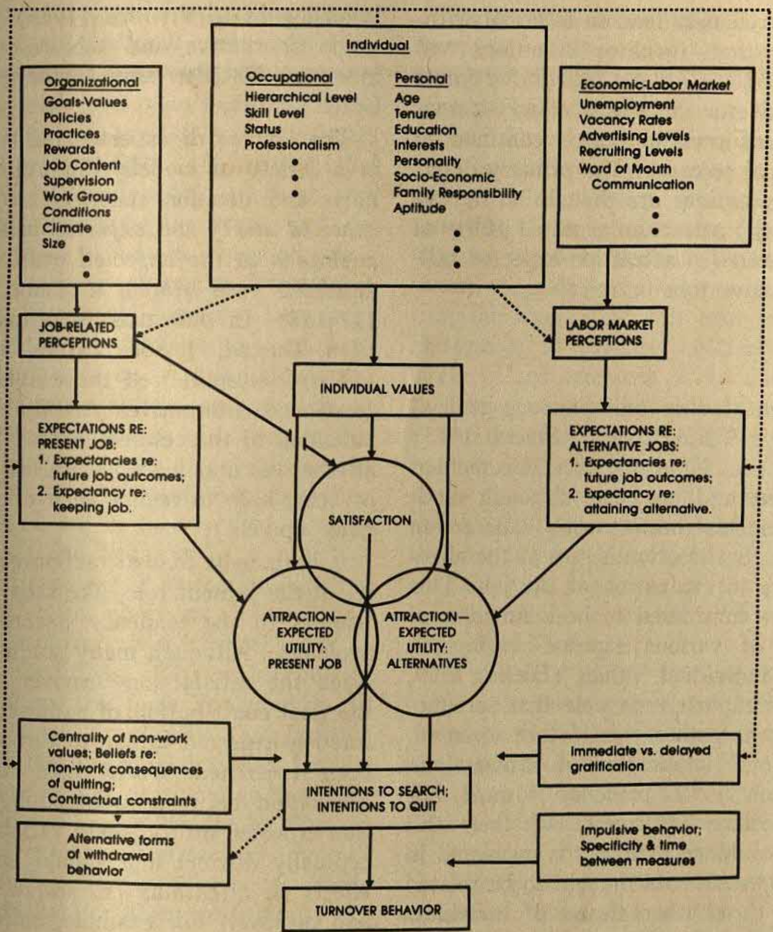

Figure 1. A schematic representation of the primary variables and process of employee turnover.

Intentions

The immediate precursor of behavior is thought to be intentions (Dulaney, 1961, 1968; Fishbein & Ajzen, 1975; Locke, 1968; Mobley, 1977; Ryan, 1970). Therefore, the best predictor of turnover should be intention to quit (see, e.g., Kraut, 1975; Mobley et al., 1978; Newman, 1974; Waters et al., 1976). The relationship between turnover and intention should be stronger the more specific the intention statement and the closer in time the measurement of the intention and the behavior. Impulsive behavior attenuates the intention–behavior relationship.

Graen and Ginsburgh (1977) observed that the more specific and closer to the act the intention measure comes, the more trivial the prediction. Although probably valid, this observation should not be interpreted as indicating that understanding the turnover process is not facilitated by including intentions and evaluating their precursors. Neither should the personnel-planning utility of assessing even more distant intentions (see, e.g., Kraut, 1975) be overlooked.

There are at least two intentions of interest, intention to search and intention to quit. Mobley (1977) suggested that intention to search and search behavior should generally precede intention to quit and turnover. Exceptions include impulsive behavior and nonsolicited attractive alternatives. Lack of perceived attractive alternatives or an un-

successful search may lead to forms of withdrawal other than turnover intentions and behavior. Relations among alternative forms of withdrawal and the effects of no alternative or an unsuccessful search continue to need additional research. The primary determinants of intentions are thought to be (a) satisfaction, (b) attraction expected utility of present job, and (c) attraction expected utility of alternative jobs or roles.

Satisfaction

The present review and previous reviews (e.g., Locke, 1976; Porter & Steers, 1973; Herzberg et al., Note 1) have documented the consistent and negative, although moderate, relationship between job satisfaction and turnover. Satisfaction is seen as the affective response to evaluation of the job. This evaluation is considered to be a function of perceptions of various aspects of the job relative to individual values (Locke, 1969, 1976). It is important to note that satisfaction is present rather than future oriented. The behavioral implication of satisfaction–dissatisfaction is a tendency toward approach–avoidance. However, whether this approach–avoidance tendency is expressed in the form of turnover is thought to be related to at least three other classes of variables: attraction expected utility of the present role, attraction expected utility of attainable alternative roles, and centrality of work values, beliefs regarding nonwork consequences of quitting–staying, and contractual constraints. Failure to consider these classes of variables may explain the absence of a stronger relationship between job satisfaction and turnover.

Attraction and Expected Utility of Present Job

Whereas satisfaction is present oriented, attraction is considered to be future oriented. Attraction is seen as being based on the expectancies that the job will lead to future attainment of various positively and negatively valued outcomes. When combined with the expectancy of being able to retain the present job, an index can be generated that is analogous to Vroom's (1964) *force* for a single alternative and the *expected utility* index of Dachler and Mobley (1973) or Graen (1976).

The concept of expected utility is applied in a variety of models. For example, economists and decision theorists frequently use *expected utility* and *expected value,* concepts analogous to the expected utility previously described (see March & Simon, 1958, pp. 137–138). In one interdisciplinary analysis, Blau, Gustad, Jessor, Parnes, and Wilcox (1956) conceptualized the evaluation of occupational alternatives as "the individual's valuation of the rewards offered by different alternatives and his appraisal of his chances of being able to realize each of the alternatives" (p. 533).

It is thought that attraction expected utility of the present role, like satisfaction, contributes to the tendency toward approach–avoidance. Although many studies have analyzed the satisfaction–turnover relationship, the dual contribution of satisfaction and attraction expected utility to turnover has not been researched. While there may be some correlation between satisfaction and attraction expected utility, these variables are conceptually distinct and should have separate effects on intentions (to search or to quit) and turnover. For example, individuals may be satisfied (or dissatisfied) with their present job, but may expect the present job to be relevant (or irrelevant) to their subsequent career. Graen and Ginsburgh (1977) found the latter belief to be significantly related to resignation. On the other hand, one may be dissatisfied with one's work group but be attracted to it because of expectations that it will facilitate the future attainment of valued outcomes or goals. The above relationships can easily be extended to include a variety of other job factors, for example, supervision, benefits, and so forth.

Just as there are multiple dimensions of satisfaction, there are multiple dimensions of attraction. The salience of these different dimensions is a function of individual differences in values. The values may be related to occupation, position level, age, tenure, and other personal variables.

One dimension of satisfaction and attraction concerns organizational goals and values. The congruence between individual and organizational goals and values has been defined by Porter et al. (1974) and Steers (1977) as one component of organizational commitment. As noted earlier, these authors also included willingness to exert effort, desire and intention to remain in the organization, and job satisfaction in the definition of commitment. The model suggested here seeks to subdivide the complex variable of commitment. Congruence between individual and organizational goals and values may be an important variable, but can be seen as distinct from, and only one of several focuses of, both satisfaction (present) and attraction (future), which are in turn related to turnover intentions and behavior.

If both satisfaction and attraction expected utility contribute to intentions to search and quit, and to turnover, then it is necessary to analyze the conditions under which one or the other makes the most contribution to variance explained. It may be that individual-level variables such as the need for immediate versus delayed gratification (see, e.g., Mischel, 1976) will aid in predicting whether satisfaction (present) or attraction (future) is most strongly related to turnover intentions and behavior for given individuals.

Attraction and Expected Utility of Alternatives

Considering both satisfaction and attraction expected utility should increase our understanding and prediction of turnover intentions and behavior. However, the attraction and attainability of alternative jobs or roles must also be considered. March and Simon (1958) presented an organizational participation model that gave a prominent role to visibility of alternatives. The March and Simon components of "perceived desirability of leaving the organization" and "perceived ease of movement from the organization" (p. 93) roughly correspond to "expectancies re: future job outcomes" and "expectancy re: attaining alternative" in Figure 1. As noted earlier, these variables have received little attention in turnover research.

The present model suggests that it is not merely the visibility of alternatives but the attraction of alternatives and the expectancy of attaining the alternatives that are most salient. Forrest et al. (1977), Mobley (1977), and Schneider (1976) are among recent authors who make a strong argument for inclusion of the variable *attraction of alternatives* in turnover research. Attraction of alternatives is defined in terms of expectations that the alternative will lead to the future attainment of various positively and negatively valued outcomes. When combined with the expectancy of being able to attain the alternative, an index can be generated that is analogous to Vroom's (1964) force for a single alternative and to the expected utility index of Dachler and Mobley (1973) or Graen (1976). (See Mobley, 1977, for a microanalytic treatment of the possible role of alternatives in search and turnover intentions and behavior.)

As noted in Figure 1, there may well be some covariation among satisfaction, attraction expected utility of the present job, and attraction expected utility of alternatives. This is to be expected, since values are common to all three and the presence or absence of attractive alternatives may result in the revaluation of one's satisfaction with or the attraction of the present role.

Moderating Variables

Although satisfaction, attraction expected utility of the present job, and attraction expected utility of alternatives are considered to be the primary determinants of turnover intentions and behavior, several other variables can be expected to moderate these relationships. To the extent that nonwork values and interests are not central to an individual's life values and interests (Dubin et al., 1975) and to the extent that an individual associates significant nonwork consequences with quitting (see, e.g., Newman, 1974), the relationships among satisfaction, attraction, and turnover intentions and behavior will be attenuated. Additionally, to the extent that an individual is bound by a contract, as for example in professional sports, the military, and certain professions,

the relationships will be attenuated during the term of the contract. Under such circumstances, it can be hypothesized that the individual who is dissatisfied, perceives little attraction in the present job, or perceives an attractive alternative may engage in other forms of avoidance and withdrawal behavior.

These suggested moderating influences, especially nonwork values and interests and nonwork consequences of turnover behavior, call attention to the need to look beyond the work setting for a complete understanding of the psychology of the turnover process.

Antecedents

The antecedents of satisfaction and attraction are considered to be organizational variables as perceived by the individual, economic variables related to availability of alternatives as perceived by the individual, and individual-level occupational and personal variables as they influence individual values, perceptions, and expectations. Although a detailed analysis of these antecedents is beyond the scope of this article, it is important to emphasize that the influence of various organizational, economic or labor market, occupational, and personal variables is through individual perceptions, expectations, and values.

Research Implications

The model described here indicates the need for multivariate research on the turnover process. As noted in the review section of this article, although the negative relationships between both age and tenure and turnover are well established, the amount of variance explained is low and the psychology of the relationships is not well understood. The model proposed here suggests that multivariate research that concurrently assesses values, job-related perceptions, external perceptions, and the previously mentioned moderating variables should facilitate an understanding of the relationship of age and tenure to turnover.

Similarly, multivariate research that concurrently assesses individual-level occupational and personal variables, job-related perceptions, external perceptions, individual values, and potential moderating variables provides a framework for integrating and understanding, at the individual level, the aggregate-level effects of various organizational and economic or labor market variables summarized by Price (1977).

Graen (1976), among others, noted that neither individuals nor organizations are fixed or static; neither is the economy or labor market. The clear implication is that understanding the turnover process will require longitudinal as well as multivariate research. Longitudinal research, not simply in terms of the collection of criterion data over time but also in terms of repeated measures of the independent variables, as recently exemplified by Graen and Ginsburgh (1977) and Porter et al. (1976), is needed.

Conclusions

In 1973 Porter and Steers observed that the then existing body of research left much to be understood about the psychology of the employee withdrawal process. Review of the subsequent research leads to a similar observation. The conceptual model suggested here calls attention to the possible main effects of satisfaction (present oriented), attraction expected utility of the current role (future oriented), and attraction expected utility of alternative roles. A number of moderating variables and constraints were suggested. The need for integrative, multivariate longitudinal research is evident if significant progress is to be made in understanding the psychology of the employee turnover process.

Reference Notes

1. Herzberg, F., Mausner, B., Peterson, R. O., & Capwell, R. F. *Job attitudes: Review of research and opinions*. Pittsburgh, Pa.: Pittsburgh Psychological Services, 1957.
2. Mangione, T. W. Turnover—Some psychological and demographic correlates. In R. P. Quinn & T. W. Mangione (Eds.), *The 1969–1970 survey of working conditions*. Ann Arbor: University of Michigan, Survey Research Center, 1973.
3. Ilgen, D. R., & Dugoni, B. L. *Initial orientation to the organization: Its impact on psychological*

processes associated with the adjustment of new employees. Kissimee, Fla.: Academy of Management, August 1977.

References

Armknecht, P. A., & Early, J. F. Quits in manufacturing: A study of their causes. *Monthly Labor Review,* 1972, *95,* 31–37.

Blau, P. M., Gustad, J. W., Jessor, R., Parnes, H. S., & Wilcox, R. C. Occupational choice: A conceptual framework. *Industrial and Labor Relations Review,* 1956, *8,* 531–543.

Brayfield, A. H., & Crockett, W. H. Employee attitudes and employee performance. *Psychological Bulletin,* 1955, *52,* 396–424.

Burke, R. J., and Wilcox, D. S. Absenteeism and turnover among female telephone operators. *Personnel Psychology,* 1972, *25,* 639–648.

Cascio, W. F. Turnover, biographical data, and fair employment practice. *Journal of Applied Psychology,* 1976, *61,* 576–580.

Dachler, H. P., & Mobley, W. H. Construct validation of an instrumentality–expectancy–task-goal model of work motivation. Some theoretical boundary conditions. *Journal of Applied Psychology,* 1973, *58,* 397–418.

Dansereau, F., Jr., Cashman, J., & Graen, G. Expectancy as a moderator of the relationship between job attitudes and turnover. *Journal of Applied Psychology,* 1974, *59,* 228–229.

Dubin, R., Champoux, J., & Porter, L. Central life interests and organizational commitment of blue collar and clerical workers. *Administrative Science Quarterly,* 1975, *20,* 411–421.

Dulaney, D. E., Jr. Hypotheses and habits in verbal "operant conditioning." *Journal of Abnormal and Social Psychology,* 1961, *63,* 251–263.

Dulaney, D. E. Awareness, rules, and prepositional control: A confrontation with S-R behavior theory. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Dunnette, M. D., Arvey, R. D., & Banas, P. A. Why do they leave? *Personnel,* 1973, *50,* 25–38.

Ekpo-Ufot, A. Self-perceived abilities relevant in the task (SPART): A potential predictor of labor turnover in an industrial work setting. *Personnel Psychology,* 1976, *29,* 405–416.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. Effect of work sample test upon self-selection and turnover of job applicants. *Journal of Applied Psychology,* 1973, *58,* 283–285.

Federico, J. M., Federico, P., & Lundquist, G. W. Predicting women's turnover as a function of extent of met salary expectations and biodemographic data. *Personnel Psychology,* 1976, *29,* 559–566.

Fishbein, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement.* New York: Wiley, 1967.

Fishbein, M., & Ajzen, I. *Belief, attitudes, intention and behavior.* Reading, Mass.: Addison-Wesley, 1975.

Fleishman, E. A., & Harris, E. F. Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology,* 1962, *15,* 43–56.

Forrest, C. R., Cummings, L. L., & Johnson, A. C. Organizational participation: A critique and model. *Academy of Management Review,* 1977, *2,* 586–601.

Gillet, B., & Schwab, D. P. Convergent and discriminant validities of corresponding Job Descriptive Index and Minnesota Satisfaction Questionnaire scales. *Journal of Applied Psychology,* 1975, *60,* 313–317.

Goodman, P. S., Salipante, P., & Paransky, H. Hiring, training, and retraining the hard-core unemployed: A selected review. *Journal of Applied Psychology,* 1973, *58,* 23–33.

Graen, G. Role-making processes within complex organizations. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology,* Chicago: Rand McNally, 1976.

Graen, G., & Ginsburgh, S. Job resignation as a function of role orientation and leader acceptance: A longitudinal investigation of organization assimilation. *Organizational Behavior and Human Performance,* 1977, *19,* 1–17.

Graen, G. B., Orris, J. B., & Johnson, T. W. Role assimilation processes in a complex organization. *Journal of Vocational Behavior,* 1973, *3,* 395–420.

Hellriegel, D., & White, G. E. Turnover of professionals in public accounting: A comparative analysis. *Personnel Psychology,* 1973, *26,* 239–249.

Herman, J. B., Dunham, R. B., & Hulin, C. L. Organizational structure, demographic characteristics and employee responses. *Organizational Behavior and Human Performance,* 1975, *13,* 206–232.

Herman, J. B., & Hulin, C. L. Studying organizational attitudes from individual and organizational frames of reference. *Organizational Behavior and Human Performance,* 1972, *8,* 84–108.

Hill, J. M., & Trist, E. L. Changes in accidents and other absences with length of service: A further study of their incidence and relation to each other in an iron and steel works. *Human Relations,* 1955, *8,* 121–152.

Hines, G. H. Achievement motivation, occupations, and labor turnover in New Zealand. *Journal of Applied Psychology,* 1973, *58,* 313–317.

Ilgen, D. R., & Seely, W. Realistic expectations as an aid in reducing voluntary resignations. *Journal of Applied Psychology,* 1974, *59,* 452–455.

Karp, H. B., & Nickson, J. W., Jr. Motivator-hygiene deprivation as a predictor of job turnover. *Personnel Psychology,* 1973, *26,* 377–384.

Koch, J. L., & Steers, R. M. Job attachment, satisfaction, and turnover among public sector employees. *Journal of Vocational Behavior,* 1978, *12,* 119–128.

Kraut, A. I. Predicting turnover of employees from measured job attitudes. *Organizational Behavior and Human Performance,* 1975, *13,* 233–243.

Lee, R., & Booth, J. M. A utility analysis of a weighted application blank designed to predict turnover for clerical employees. *Journal of Applied Psychology,* 1974, *59,* 516–518.

Lefkowitz, J., & Katz, M. L. Validity of exit interviews. *Personnel Psychology,* 1969, *22,* 445–455.

Locke, E. A. Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance,* 1968, *3,* 157–189.

Locke, E. A. What is job satisfaction? *Organizational Behavior and Human Performance,* 1969, *4,* 309–336.

Locke, E. A. The nature and consequences of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand-McNally, 1976.

March, J. G., & Simon, H. A. *Organizations.* New York: Wiley, 1958.

Marsh, R., & Mannari, H. Organizational commitment and turnover: A predictive study. *Administrative Science Quarterly,* 1977, *22,* 57–75.

Mirvis, P. H., & Lawler, E. E., III. Measuring the financial impact of employee attitudes. *Journal of Applied Psychology,* 1977, *62,* 1–8.

Mischel, W. *Introduction to personality.* New York: Holt, Rinehart & Winston, 1976.

Mobley, W. H. Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology,* 1977, *62,* 237–240.

Mobley, W. H., Horner, S. O., & Hollingsworth, A. T. An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology,* 1978, *63,* 408–414.

Newman, J. E. Predicting absenteeism and turnover: A field comparison of Fishbein's model and traditional job attitude measures. *Journal of Applied Psychology,* 1974, *59,* 610–615.

Pettman, B. D. Some factors influencing labour turnover: A review of the literature. *Industrial Relations Journal,* 1973, *4,* 43–61.

Porter, L. W., Crampon, W. J., & Smith, F. J. Organizational commitment and managerial turnover: A longitudinal study. *Organizational Behavior and Human Performance,* 1976, *15,* 87–98.

Porter, L. W., & Steers, R. M. Organizational, work, and personal factors in employee turnover and absenteeism. *Psychological Bulletin,* 1973, *80,* 151–176.

Porter, L. W., Steers, R. M., Mowday, R. T., & Boulian, P. V. Organizational commitment, job satisfaction, and turnover among psychiatric technicians. *Journal of Applied Psychology,* 1974, *59,* 603–609.

Price, J. L. The measurement of turnover. *Industrial Relations Journal,* 1975–1976, *6,* 33–46.

Price, J. L. *The study of turnover.* Ames: Iowa State University Press, 1977.

Rice, A. K., & Trist, E. L. Institutional and subinstitutional determinants of change in labor turnover. *Human Relations,* 1952, *5,* 347–372.

Robinson, W. S. Ecological correlations and the behavior of individuals. *American Sociological Review,* 1950, *15,* 351–357.

Ryan, T. A. *Intentional behavior: An approach to human motivation.* New York: Ronald Press, 1970.

Schneider, J. The "greener grass" phenomenon: Differential effects of a work context alternative on organizational participation and withdrawal intentions. *Organizational Behavior and Human Performance,* 1976, *16,* 308–333.

Schwab, D. P., & Oliver, R. L. Predicting tenure with biographical data: Exhuming buried evidence. *Personnel Psychology,* 1974, *27,* 125–128.

Smith, P. C., Kendall, L. M., & Hulin, C. L. *The measurement of satisfaction in work and retirement.* Chicago: Rand McNally, 1969.

Steers, R. M. Antecedents and outcomes of organizational commitment. *Administrative Science Quarterly,* 1977, *22,* 46–56.

Vroom, V. H. *Work and motivation.* New York: Wiley, 1964.

Wanous, J. P. Effects of a realistic job preview on job acceptance, job attitudes, and job survival. *Journal of Applied Psychology,* 1973, *58,* 327–332.

Waters, L. K., & Roach, D. Job attitudes as predictors of termination and absenteeism: Consistency over time and across organizational units. *Journal of Applied Psychology,* 1973, *57,* 341–342.

Waters, L. K., Roach, D., & Waters, C. W. Estimate of future tenure, satisfaction, and biographical variables as predictors of termination. *Personnel Psychology,* 1976, *29,* 57–60.

Woodward, N. The economic causes of labour turnover: A case study. *Industrial Relations Journal,* 1975–1976, *6,* 19–32.

Received January 12, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 3, 523–548

Associative and Nonassociative Theories of the UCS Preexposure Phenomenon: Implications for Pavlovian Conditioning

Alan Randich and Vincent M. LoLordo
Dalhousie University, Halifax, Canada

Excitatory Pavlovian conditioning of a discrete conditioned stimulus is attenuated by prior exposure to the unconditioned stimulus. The unconditioned stimulus preexposure phenomenon is observed in a variety of Pavlovian conditioning procedures as diverse as eyelid conditioning, the conditioned emotional response, and conditioned taste aversion learning. This article discusses the variables that affect the unconditioned stimulus preexposure phenomenon and uses this information in evaluating both associative and nonassociative accounts of the phenomenon. At least one associative account, based on context blocking, and at least one nonassociative account, based on central habituation of the emotional response to the unconditioned stimulus, remain viable.

The primary goals of research in Pavlovian conditioning are to determine the variables that influence the formation of conditioned responses and then to specify their mechanisms of action. A number of investigators are currently examining how one such variable, the organism's experience with the unconditioned stimulus, affects the course of conditioning. It has been shown in a variety of Pavlovian conditioning paradigms that exposure to the unconditioned stimulus (UCS) prior to the initiation of pairings of the conditioned stimulus (CS) and the UCS attenuates the formation of the excitatory conditioned response (CR). The UCS preexposure effect interests many investigators because of their conviction that a thorough analysis of this phenomenon will further our understanding of the necessary conditions for Pavlovian conditioning.

In general, the UCS preexposure effect has been viewed from two different perspectives, each of which bears on the possible mechanism(s) of Pavlovian conditioning. Associative accounts of the UCS preexposure effect argue that during the preexposure phase, an animal learns some relation between the occurrence of the UCS and other events that occur in the situation, such as the relation between the UCS and contextual stimuli. The learning that occurs in the preexposure phase somehow interferes with the formation of an association between a nominal CS and the UCS in the second phase of the experiment. For example, a context-blocking interpretation of the UCS preexposure phenomenon (e.g., Rescorla & Wagner, 1972) posits that some stimulus aspect (X) of the experimental situation acquires associative strength during preexposure to an aversive UCS. Conditioning of Stimulus X reduces the amount of associative strength that a nominal stimulus (A) can acquire during subsequent excitatory conditioning in that environment (i.e., pairings of AX with the UCS). As a result, the rate of acquisition of an excitatory CR to Stimulus A is attenuated. On the other hand, nonassociative accounts generally hold that preexposure to the UCS reduces the organism's reactivity to subsequent applications of that UCS. Such a reduction in reactivity may result from either central habituation or peripheral sensory adaptation of the re-

This research was supported by Grant A-9585 from the National Research Council of Canada. We wish to thank J. Willner and A. L. Riley for comments on an earlier version of the article and J. Lord for preparation of the manuscript.

Requests for reprints should be sent to Alan Randich, who is now at the Department of Psychology, Yale University, Box 11A, Yale Station, New Haven, Connecticut 06520.

sponse to the UCS. In any case, reduced reactivity to the UCS should result in attenuated excitatory conditioning to a CS paired with that UCS.

The present article reviews the effects of preconditioning exposure to the UCS on the acquisition of excitatory CRs and evaluates the various hypotheses that have been advanced to explain this phenomenon. This review focuses on Pavlovian conditioning using aversive UCSs, since most of the published experiments on this phenomenon have used such stimuli. Moreover, the review focuses only on what Domjan and Best (1977) referred to as the durable UCS preexposure effect, which occurs even when long delays intervene between UCS preexposure and conditioning. The recent demonstrations of a transient or proximal UCS preexposure effect, which occurs only when the UCS is preexposed a short time before the conditioning trial, are not discussed (cf. Domjan & Best, 1977; Terry, 1976; Wagner, 1976).

Preconditioning Exposure to the UCS

In a typical Pavlovian conditioning experiment, a neutral stimulus (CS) is presented prior to and terminates with the presentation of a UCS. After repeated pairings of the CS and UCS, the previously neutral stimulus comes to elicit a CR. In this section, studies that present the UCS prior to the initiation of CS–UCS pairings, or preexpose the UCS, are organized according to the nature of their experimental paradigm. The paradigms include human and rabbit eyelid conditioning procedures, conditioned emotional response (CER) procedures, and conditioned taste aversion procedures.

Human and Rabbit Eyelid Conditioning

Taylor (1956) preexposed each of three groups of human subjects to 50 presentations of an air puff UCS to the cornea of the eye. The intensity of the UCS was either 15 mm, 30 mm, or 80 mm. In each group the amplitude of the unconditioned eyeblink response (UCR) decreased significantly with repeated presentations of the UCS, but there were no significant between-groups differences in the amplitude of the UCR. Each subject then received 50 presentations of a 520-msec light paired with an air puff of 30 mm to the cornea. The number of eyeblink CRs was greatest in the group that received no preexposure to the UCS, whereas the number of CRs in the groups preexposed to the UCS was an inverse function of the intensity of the UCS used during the preexposure phase.

Taylor considered the possibility that groups preexposed to the air puff UCS showed a reduced number of eyeblink CRs during excitatory conditioning because of peripheral sensory adaptation of the UCR. However, she rejected peripheral sensory adaptation as a complete explanation of these effects because there was no evidence of differential UCR magnitudes in the three UCS intensity groups at any time during the preexposure phase. Such differences would be expected to accompany differential sensory adaptation.

As an alternative explanation, Taylor suggested that a more general emotional or fear response elicited by the UCS was reduced in magnitude as a result of preexposure to the UCS. This argument requires two assumptions. It assumes that acquisition of an excitatory CR depends in part on an association formed between the CS and the organism's emotional reaction to the UCS. Konorski (1967) made a similar, albeit more general, point, asserting that in all Pavlovian aversive conditioning, a fear (preparatory) response is first conditioned to the CS and that the presence of this preparatory CR is necessary for the formation of a consummatory CR, for example, the eyeblink response. Second, the argument assumes that preexposure to the UCS reduces the magnitude of this general emotional response elicited by the UCS, even though the specific UCR (in Konorski's terms, the consummatory response) measured by an experimenter may be relatively unaffected. The slower rates of acquisition of an excitatory CR in subjects preexposed to the UCS would then reflect a reduced emotional responsiveness relative to that of subjects not preexposed to the UCS.

Kimble and Dufort (1956) also examined the effect of preexposure to an air puff UCS on subsequent acquisition of the eyeblink response in humans. One group of subjects were preexposed to 20 air puffs (90 mm), while a second group of subjects received no preexposure treatment. Both groups then received 60 conditioning trials in which a .25-sec light CS was paired with an air puff UCS (90 mm). The group preexposed to the UCS acquired the eyeblink CR slower than the no-preexposure controls. These authors also offered an explanation based on motivational factors, but in one respect it is the opposite of the account suggested by Taylor. According to their view, preexposure to the UCS produces an increase in drive, and this drive energizes the subject's dominant habit regardless of its nature. If the subject's dominant habit before conditioning were not the CR and, indeed, if the dominant response to the CS were antagonistic to the CR, then increasing the subject's drive would interfere with subsequent excitatory conditioning.

In a similar experiment, Hobson (1968) preexposed college men and women rated as high- or low-anxiety subjects on the Taylor Manifest Anxiety Scale to air puffs (1.5 lb/sq. in.) delivered to the cornea of the eye for 0, 35, or 70 trials. Following preexposure to the UCS all subjects received excitatory conditioning of the eyelid response. The number of CRs obtained during excitatory conditioning was inversely related to the number of preexposures to the UCS. In addition, high-anxiety subjects acquired the CR significantly faster than low-anxiety subjects, although both classes of subjects showed a preexposure effect. If manifest anxiety is considered a source of drive, then Kimble and Dufort's account predicts less interference and hence faster eyelid conditioning in low-anxiety subjects than in high-anxiety subjects. On the other hand, perhaps Hobson's anxiety variable represents some aspect of general emotional responsiveness. Low-anxiety subjects preexposed to the UCS may acquire the excitatory CR at an even slower rate than high-anxiety subjects preexposed to the UCS because of a reduced emotional responsiveness to aversive stimuli.

Siegel and Domjan (1971) preexposed one group of rabbits to 550 100-msec, 200-V shocks to the infraorbital region of the eye, whereas a second group of rabbits received no preexposure to the UCS. Infraorbital shock elicits an unconditioned extension of the nictitating membrane, or third eyelid, in the rabbit. The nictitating membrane response was then conditioned in both groups by pairing a 500-msec tone with a UCS of the same intensity as that used during the preexposure phase. The rabbits preexposed to the UCS showed a slower rate of acquisition of the nictitating membrane CR than the no-preexposure controls.

In a more extensive study, Mis and Moore (1973) preexposed groups of rabbits to electric shock delivered to the infraorbital region of the eye and varied both the number of preexposed UCSs (0, 50, 200, or 350) and the interval between the last preexposed UCS and subsequent excitatory conditioning of the nictitating membrane response (30 sec vs. 24 hours). There were no significant changes in the amplitude of the UCR over trials during the preexposure phase. The rate of acquisition of the nictitating membrane response was reduced in both the 30-sec-delay and the 24-hour-delay groups, relative to controls that received no preexposure, and was an inverse function of the number of shocks administered during the preexposure phase. Moreover, rabbits in the 30-sec-delay groups acquired the excitatory CR at a slower rate than their counterparts in the 24-hour-delay groups. In general, there were no differences in the asymptotic levels of excitatory conditioning in the various groups after repeated CS–UCS pairings.

In a second experiment, groups of rabbits were preexposed to (a) no shocks, (b) 450 1-mA shocks, (c) 450 3-mA shocks, or (d) 450 5-mA shocks to the infraorbital region of the eye. There were no decrements in the amplitude of the UCR over successive preexposures to the UCS, and the UCRs to shocks of various intensities were comparable. All groups then received immediate excitatory conditioning of the nictitating membrane response; the CS was paired with the 3-mA UCS. The nictitating membrane CR

was acquired fastest in the no-shock and 1-mA groups, and slowest in the 3-mA and 5-mA groups. However, all groups eventually attained approximately the same asymptotic level of conditioning.

Mis and Moore's experiments show that the decremental effect of preexposure to the UCS on the rate of acquisition of the nictitating membrane response is a direct function of the number of preexposures to the UCS (cf. Hobson, 1968), a direct function of the intensity of the prior UCSs (cf. Taylor, 1956), and an inverse function of the time interval between the last preexposure to the UCS and the initiation of excitatory conditioning. Thus, analogous parametric manipulations during preexposure to the UCS, for example, variations of the number of preexposures, produce similar effects on conditioning of the eyelid response in humans and conditioning of the nictitating membrane response in rabbits.

It is unlikely that sensory adaptation to the UCS can account for the retarded acquisition of excitatory conditioning in Mis and Moore's experiments because the amplitude of the UCR did not change during preexposure to the UCS. Furthermore, different UCS intensities produced comparable UCRs, as was the case in Taylor's study of the human eyeblink.

Several investigators (e.g., Rudy, Iwens, & Best, 1977; Tomie, 1976) have asserted that the UCS preexposure effect can be explained in terms of blocking of conditioning to the nominal CS in the second phase of the experiment as a result of prior excitatory conditioning of some cues that were present in both phases (Kamin, 1969). In Kamin's original demonstration of blocking, one group of rats were exposed to repeated presentations of Stimulus A paired with electric shock in a CER paradigm (Estes & Skinner, 1941), whereas a second group of rats received no treatment. Then both groups received presentations of a compound stimulus (AB) paired with an electric shock of the same intensity as that used during the first phase of the experiment. Following the compound conditioning, the conditioned properties of A and B were assessed. The group that had received no treatment in the initial phase showed strong conditioned suppression during both Stimulus A and Stimulus B. The group that had received conditioning of Stimulus A in the initial phase showed strong conditioned suppression during test presentations of Stimulus A alone but little or no suppression during presentations of Stimulus B alone. Thus, prior conditioning of Stimulus A is said to block the conditioning of Stimulus B that normally occurs when the compound stimulus, AB, is paired with a UCS.

A blocking interpretation of Mis and Moore's UCS preexposure effect would hold that some static cues present in the experimental environment during the preexposure phase, call them contextual cues, acquire associative strength by virtue of their presence when the UCS occurs. Then when the excitatory conditioning phase begins, the nominal CS conveys less information about the UCS than in control groups because contextual cues already predict the occurrence of the UCS. Hence, acquisition of the excitatory CR is retarded.

Mis and Moore concluded that a blocking interpretation of their data is highly speculative because there is no independent evidence that the rabbit nictitating membrane response can be conditioned to background cues (cf. Plotkin & Oakley, 1975; Moore, Note 1). However, this would not preclude conditioning of some preparatory, emotional response to the background stimuli. Suppose that the nictitating membrane CR can only be evoked by CSs that already evoke an increase in a preparatory, emotional response. Then prior conditioning of the emotional response to contextual cues could block conditioning of the emotional response to the nominal CS in the second phase, thereby blocking conditioning of the nictitating membrane response. However, if one assumes that the background stimuli are much less salient stimuli than the nominal CS, then one should expect a relatively small blocking effect (Feldman, 1975; Hall, Mackintosh, Goodall, & dal Martello, 1977).

Mis and Moore's interpretation of their results was that preexposure to the UCS made the rabbits less emotionally reactive

to the UCS and that reduced responsiveness to the UCS slowed the rate of excitatory conditioning—an argument identical to that proposed by Taylor (1956).

Conditioned Emotional Response Paradigms

Kamin (1961) examined the effect of prior exposure to electric shock on the subsequent acquisition of a conditioned emotional response (CER). In Experiment 1, two groups of rats were trained to bar press for food reinforcement on a variable-interval (VI) schedule. One group of rats then received 10 days of exposure to a .5-sec, .8-mA electric shock delivered through a grid floor at the rate of four shocks per day. A second group of rats received no preexposure treatment, but both groups of rats continued bar pressing for food reinforcement during this period. All rats then received CER training in which pairings of a 3-min white noise CS and a .5-sec, .8-mA electric shock were superimposed on the VI baseline. The group preexposed to the UCS showed a retarded rate of acquisition of conditioned suppression relative to the no-treatment controls.

Kamin replicated these findings in a second experiment using a 1.0-mA electric shock in both the preexposure and the conditioning phases. However, he observed that the decremental effect of preexposure to the UCS was reduced relative to the first experiment. Further, Kamin commented that a group of rats preexposed to a .72-mA electric shock readily acquired the CER if the intensity of the UCS used in the excitatory conditioning phase was 2.4 mA.

In a third experiment, Kamin assessed the effect of UCS intensity. Essentially the same procedure was used as in Experiment 1, except that three groups of rats were preexposed to shock intensities of .28 mA, .49 mA, or .85 mA. All groups then received CER conditioning with a .85-mA UCS. Groups that were preexposed to the UCS showed a retarded rate of acquisition of the CER relative to no-preexposure controls, and the magnitude of the retardation effect was a direct function of the intensity of the UCS used during preexposure (cf. Mis & Moore, 1973; Taylor, 1956).

A final experiment was conducted to determine whether the delivery of shocks during the preexposure phase retarded the acquisition of a CER because of contiguous pairings of the bar-press response and the UCS. One group of rats were preexposed to shock with the bar present and the VI contingency in effect (on-baseline treatment), whereas a second group of rats were preexposed to shock with the bar absent and the VI contingency eliminated (off-baseline treatment). Both groups showed a slower rate of acquisition of the CER than did the no-preexposure controls, but the rats that received shocks in the off-baseline treatment acquired the CER faster than those that received the on-baseline treatment. Kamin reported that on- versus off-baseline preexposure treatments did not produce differences in the baseline rates of responding on the VI schedule during excitatory conditioning.

Kamin suggested that some central habituation of emotional reactivity to the UCS resulted from preexposure to the UCS and produced retarded acquisition of the CER. This account is similar to the general emotional reactivity account of the preexposure phenomenon suggested by Taylor and by Mis and Moore. Note that such an explanation can also account for the differences between the on- versus off-baseline preexposure effects. Suppose that central habituation of emotional reactivity is relatively specific to the situation in which shock is administered. Then one would expect more stimulus generalization decrement of habituation, and hence faster excitatory conditioning, when the stimulus conditions are different in the preexposure and conditioning phases.

Kamin's experiments extend the UCS preexposure phenomenon to a new species, rats, and to another widely studied Pavlovian conditioning paradigm, CER. Further, they suggest that the magnitude of the retardation of the acquisition of a CER is a direct function of the intensity of the UCS used during preexposure. However, Kamin included no group preexposed to a UCS intensity that exceeded the intensity used in the subse-

quent excitatory conditioning phase. Thus, it is possible that the decremental effect of preexposure to a UCS on the acquisition of a CER simply reflects similarities between the intensity of the preexposure UCS and the intensity of the UCS used during excitatory conditioning, rather than an effect of UCS intensity during preexposure per se. Such an account suggests that the magnitude of the UCS preexposure effect would also be diminished when the UCS is much more intense during preexposure than during excitatory conditioning. Indeed, Randich (1978) obtained this outcome using electric shock as a UCS and the CER as a measure of excitatory conditioning. Groups of rats were preexposed to 0-mA, .3-mA, .5-mA, .8-mA, or 1.3-mA unsignaled electric shocks for 10 days. All groups then received pairings of a 3-min white noise CS with a .8-mA electric shock. All groups preexposed to electric shock showed retarded acquisition of the CER relative to the no-preexposure control group (0 mA), but the greatest attenuation of CER acquisition occurred in the group both preexposed and conditioned with the .8-mA electric shock. The groups that received a shift in UCS intensity between the preexposure and CER conditioning phases acquired the CER faster than did the nonshifted group.

Thus, if groups of subjects are preexposed to a wide range of UCS intensities and all are conditioned at an intermediate UCS intensity, then the magnitude of the UCS preexposure effect is a direct function of the intensity of the preexposed UCS in both human eyelid (Taylor, 1956) and rabbit nictitating membrane (Mis & Moore, 1973) paradigms, but is an inverted-U-shaped function of the intensity of the preexposed UCS in the CER paradigm (Randich, 1978). The different functions yielded by different response systems have important implications for understanding the mechanism(s) of the UCS preexposure phenomenon.

For example, a context-blocking account of the UCS preexposure phenomenon, at least in the interpretation offered by the model of Rescorla and Wagner (1972), predicts that the magnitude of the UCS preexposure phenomenon should be an increasing function of the intensity of the UCS used during preexposure. However, such a view would have difficulty explaining the inverted-U function obtained by Randich, unless one assumes that shifts in UCS intensity between the preexposure and conditioning phases result in stimulus generalization decrement of the associative strength of contextual cues and thus in a decrease in context blocking. Granting that possibility, however, why would stimulus generalization decrement not occur in the eyelid and nictitating membrane procedures?

Brimer and Kamin (1963), Kremer (1971), and Baker (1974) have also reported similar attenuation of a CER following off-baseline preexposure to shocks. On the other hand, Mackintosh (1973) failed to observe a significant preexposure effect in a CER procedure, although preexposed rats on the average acquired the CER slower than no-preexposure controls (see also, Pearce & Dickinson, 1975, Experiment 1).

Rescorla (1973) preexposed one group of rats to 72 2-sec loud noise presentations (112 dB; SPL), whereas a second group of rats received no preexposure treatment. Both groups then received pairings of a 30-sec visual stimulus with the loud noise UCS in a CER procedure. The group preexposed to the loud noise UCS acquired conditioned suppression at the same rate as no-preexposure controls, that is, there was no preexposure effect on acquisition of the CER. However, during a 3-day extinction procedure, in which the CS was presented alone, the CER of the group preexposed to loud noise extinguished more rapidly than did the CER of the no-preexposure controls. It is difficult to explain the failure of this preexposure treatment to retard the formation of a CER. Further investigations of this discrepancy should focus on the distal nature of the UCS, the nature of the UCR to loud noise (cf. Bolles & Seelbach, 1964), and the parametric effects of intensity with a loud noise UCS.

In another experiment, Rescorla (1974) preexposed three groups of rats to a total of eight .5-sec shocks of .5-mA, 1.0-mA, or 3.0-

mA intensity. Each preexposed shock was preceded by a 2-min flashing light in an effort to prevent fear conditioning to contextual stimuli. Following a 72-hour delay in which the food-reinforced baseline response was reestablished, a 2-min tone CS was repeatedly paired with a .5-sec, .5-mA electric shock. Rescorla found that the rate at which the CER was acquired was a direct function of the intensity of the UCS used during the preexposure phase. However, a control group that received no preexposure treatment was not included in this experiment, making it impossible to determine whether the groups preexposed to a UCS showed an accelerated or retarded rate of acquisition of the CER.

Randich (1978) provided a more complete investigation of the effects of preexposure to a signaled UCS on acquisition of a CER by including a no-preexposure control group as well as a group preexposed to a weaker shock than the shock used during CER conditioning. Groups of rats were preexposed to a total of 30 .5-sec shocks of .5-mA, .8-mA, or 1.3-mA intensity, distributed over 10 days. Each preexposed shock was preceded by a 3-min light stimulus, and a no-preexposure control group (0 mA) received exposure to the light stimulus alone. All groups then received pairings of a 3-min white noise CS with the .8-mA electric shock. All groups preexposed to electric shock acquired the CER at a slower rate than the group preexposed to the light stimulus alone, and the greatest attenuation of CER conditioning occurred in the group both preexposed and conditioned with the .8-mA electric shock. This pattern of results includes a replication of those obtained by Rescorla (1974). Moreover, the pattern of results obtained with preexposure to signaled shocks duplicates the pattern of results obtained by Randich (1978) with preexposure to unsignaled shocks. Since one would assume that signaling the preexposed UCS would minimize conditioning of contextual cues, the similar pattern of results obtained with preexposure to signaled versus unsignaled electric shock suggests that, at a minimum, context blocking is not the only factor responsible for the UCS preexposure phenomenon with a CER procedure.

Conditioned Taste Aversion Paradigms Within UCS Comparisons

Elkins (1974) preexposed groups of rats to injections of the emetic cyclophosphamide (12.5 mg/kg) for 0, 1, 3, or 6 days over the course of a 2-week period. All subjects were then deprived of water. They were then presented with a .1% saccharin solution for a minimum of 10 minutes and for 5 minutes after the onset of drinking. Five minutes after the removal of saccharin, each group was injected with cyclophosphamide (12.5 mg/kg). On the day following the conditioning trial, and for 60 days thereafter, each subject was given free access to both saccharin solution and tap water presented in separate bottles, that is, extinction testing. In general, Elkins found the initial attenuation of the aversion to saccharin to be a direct function of the number of preexposures to cyclophosphamide injections. Moreover, the rate at which the aversion to saccharin extinguished over the 60-day period was a direct function of the number of preexposures to the UCS.

In a similar design, Vogel (Note 2) showed that preexposure to amobarbital attenuated the aversion induced by pairings of a novel taste with amobarbital as a direct function of the number of preexposed UCSs (0, 1, 3, or 5).

Cannon, Berman, Baker, and Atkinson (1975) preexposed three groups of rats to ethanol intubation (4 g/kg) for 1, 3, or 5 successive days, whereas two other groups of rats received intubation with equivalent volumes of water. All of the groups preexposed to ethanol and one group preexposed to water then received saccharin–ethanol (4 g/kg) pairings for 3 days, while the second group preexposed to water received saccharin–water pairings. Conditioning trials were spaced 3 days apart. None of the groups preexposed to ethanol showed an aversion to saccharin following the first pairing of saccharin and ethanol, that is, they drank as much saccharin solution as the group that received no preexposure treatment and a saccharin–water pairing. On the other hand, the control group given no preexposure followed by a pairing of saccharin and ethanol showed

a strong aversion to saccharin after a single conditioning trial. By the third conditioning trial, animals that had been preexposed to ethanol did show an aversion to saccharin, but it was attenuated relative to controls. The magnitude of the attenuation effect was a direct function of the number of preexposures to ethanol.

Elsmore (1972) preexposed groups of rats to delta-9-tetrahydrocannabinol (THC; 10 mg/kg) for 0, 1, 2, 4, or 8 days. All groups then received a single pairing of saccharin and delta-9-THC (10 mg/kg) followed by a two-bottle test in which they could drink either saccharin or water. All groups showed approximately the same aversion to saccharin, although there was a nonsignificant trend indicating that the aversion was weakest in the groups that received the greatest number of preexposures to delta-9-THC. It is important to note that a two-bottle test is a good method for detecting very weak aversions, but is less effective than a one-bottle test in demonstrating between-groups differences in the strength of an aversion when all groups have a moderate to strong aversion.

In summary, the data on the effect of the number of preexposures to the UCS on subsequent conditioning of a taste aversion generally complement the findings reported with human and rabbit eyelid conditioning (Hobson, 1968; Mis & Moore, 1973). The data of Cannon et al. (1975) also suggest that repeated conditioning trials may sometimes be required to reveal the effect of some parametric variation of a UCS preexposure treatment that is not detected following a single conditioning trial. Elsmore's (1972) failure to obtain a preexposure effect with delta-9-THC may reflect either the use of a single conditioning trial or, as noted above, the use of a two-bottle test.

The effect of variations in the interval between preexposure to a drug UCS and subsequent conditioning of an aversion to a taste paired with that drug has been investigated in several experiments by Cappell and LeBlanc. They showed that the magnitude of the attenuating effect of prior exposure to either D-amphetamine (Cappell & LeBlanc, 1975) or morphine (Cappell & LeBlanc, 1977) on subsequent taste aversion learning with the same UCS decreases as the time interval increases between the last preexposure to the UCS and the start of conditioning (cf. Mis & Moore, 1973). Moreover, the dissipation of the UCS preexposure effect with delays to the start of conditioning was more pronounced with amphetamine than with morphine.

Several studies have examined the effect of drug dosage during preexposure on the magnitude of the UCS preexposure effect. Cannon et al. (1975) preexposed groups of rats to a single saline injection or an injection of either .12-M or .36-M lithium chloride (LiCl; an emetic). Saccharin was then paired with injections of saline, .12-M LiCl, or .36-M LiCl in a factorial design. The decremental effect of preexposure to LiCl on subsequent taste aversion learning was a direct function of the preconditioning dosage of LiCl. The magnitude of the decremental effect of preexposure to a given dosage of LiCl was an inverse function of the conditioning dosage; that is, the greater the conditioning dosage, the greater the aversion, given a constant preconditioning dosage. Thus, there was no interaction between preconditioning dosage and conditioning dosage; these variables appear to combine in a simple linear fashion. The pattern of results obtained in this study is similar to that obtained in human eyeblink and rabbit nictitating membrane conditioning procedures but not in the CER procedure with rats.

Parker, Failor, and Weidman (1973) preexposed two groups of rats to chronic morphine treatment over a 25-day period, while a third group of rats received no preexposure treatment. The dosage of morphine was gradually increased over the preexposure period, and reached a final level of 140 mg/kg per day, a large dose. A preconditioning preference test between sucrose-octa-acetate (SOA) solution and water was then administered to all rats during a period in which the addicted rats were deprived of morphine for 96 hours. Subsequently, one group of rats that had been preexposed to morphine were repeatedly injected with morphine after consuming SOA; a second group of rats preexposed to morphine were injected with morphine after

no-liquid sessions; and a third group of rats that received no preexposure treatment were injected with morphine after consuming SOA. These injections were spaced 72 hours apart and thus occurred when the previously addicted rats were suffering withdrawal stress. The rats preexposed to morphine prior to pairings of SOA and morphine showed an attenuated aversion to SOA relative to the rats that received no preexposure treatment prior to SOA–morphine pairings. In addition, rats preexposed to morphine prior to SOA–morphine pairings showed a preference for SOA when compared with rats preexposed to morphine and then given morphine injections after no-liquid sessions. Perhaps even stronger testimony to the effect of chronic pretreatment with morphine on the impact of SOA paired with morphine is that seven rats that were not addicted to morphine but were conditioned with morphine died. No rats addicted to morphine and conditioned with morphine died.

The authors concluded that chronic preexposure to morphine creates an unnatural need state for the drug. By this view, SOA–morphine pairings do not induce an aversion in rats pretreated with morphine because SOA comes to signal alleviation of withdrawal symptoms induced by the need for morphine. Indeed, SOA that has been paired with morphine might even be expected to elicit a stronger approach response in rats pretreated with morphine than in appropriate controls.

LeBlanc and Cappell (1974) attempted to test the unnatural need state hypothesis in two experiments. In Experiment 1, chronic preexposure to either a large (200 mg/kg) or moderate (40 mg/kg) daily dose of morphine eliminated the formation of an aversion to saccharin induced by saccharin–morphine (20 mg/kg) pairings in the morning, and there was no evidence of any conditioning with repeated taste–drug pairings. However, this outcome may reflect more than a UCS preexposure effect because the groups that were preexposed to morphine received a supplemental dose of morphine each afternoon to maintain the normal daily level of morphine. For example, the group preexposed to 200 mg/kg of morphine per day and conditioned with 20 mg/kg of morphine per day received a supplement of 180 mg/kg of morphine in the afternoon, following a taste–drug pairing in the morning. The implications of providing supplemental doses of morphine are unclear because little is known about the effects of interpolated presentations of the UCS between conditioning trials. However, it is possible that since these supplemental doses of morphine constituted from 50%–90% of the animals' normal daily morphine intake, saccharin–morphine pairings signaled an event of little metabolic consequence in satisfying a need state for morphine. Alternatively, this procedure could be viewed within the context of contingency theory, according to which the administration of morphine in the afternoon degrades the positive contingency between saccharin and morphine. This should render the CS a less effective signal (cf. Rescorla, 1967), resulting in little conditioning.

In a second experiment, chronic preexposure to either a large (20 mg/kg) or small (4 mg/kg) daily dose of D-amphetamine attenuated an aversion induced by pairing saccharin with D-amphetamine (1 mg/kg). The magnitude of the decremental effect of preexposure to D-amphetamine on the aversion to saccharin was larger in the high-dependence group than in the low-dependence group (see also, Goudie, Thornton, & Wheeler, 1976).

LeBlanc and Cappell argued that since there is no convincing evidence for a need state artificially induced by D-amphetamine withdrawal, there is no need to postulate a need state to account for the drug preexposure effect, even for so-called drugs of abuse. Instead, LeBlanc and Cappell argued that drug tolerance, or a diminished responsiveness to repeated administration of a constant dose of a drug, develops to presentations of either morphine or D-amphetamine during the preexposure phase and retards the development of an aversion to a taste paired with those drugs.

Cannon, Baker, and Berman (1977) have obtained evidence that is compatible with the claim that tolerance for the effects of a

drug contributes to the drug UCS preexposure effect. Rats that were repeatedly preexposed to ethanol and then given six pairings of saccharin and ethanol showed no tendency to acquire a saccharin aversion, whereas nonpreexposed controls formed a strong aversion. On the day following the last conditioning trial, the rats were intubated with ethanol and 30 minutes later were placed on a rotarod (a test of balance). Saccharin consumption and time on the rotarod, that is, maintenance of balance, were significantly positively correlated, $r(15) = .54$, leading Cannon et al. to conclude that tolerance for the effects of ethanol contributed a drug-specific component to their UCS preexposure effect.

On the other hand, Vogel (Note 2) asserted that drug tolerance is not a necessary condition of the drug preexposure effect. One group of rats received daily injections of amobarbital (80 mg/kg) for 5 days, whereas a second group of rats received injections of the vehicle. Both groups then received sweetened milk paired with amobarbital (80 mg/kg). The rats preexposed to amobarbital showed an attenuated aversion to sweetened milk relative to the no-preexposure control group. Vogel suggested that drug tolerance cannot entirely explain these results because the duration of sleep in rats preexposed to amobarbital was not significantly reduced relative to that of rats preexposed to the vehicle. Reduced sleeping time is often used as a measure of tolerance for barbiturates. In addition, if the attenuated taste aversion in animals preexposed to amobarbital reflects drug tolerance alone, there should be a correlation between reduced sleeping times and weaker taste aversions. Vogel found that this correlation was only .41. Perhaps stronger testimony against a drug tolerance hypothesis is Vogel's demonstration that a single preexposure to amobarbital given 2 days prior to a taste–drug pairing attenuated the aversion. Vogel argued that a single preexposure to amobarbital is unlikely to produce drug tolerance, although he presented no independent evidence to show that drug tolerance did not occur.

The rats in Vogel's experiments were deprived of food and water prior to each amobarbital injection during both the preexposure phase and the conditioning phase. Vogel recognized the possibility that during preexposure an aversion may have been conditioned to drive cues associated with deprivation states and that this aversion may have blocked subsequent taste aversion learning. Vogel tested this possibility by preexposing a group of rats to a total of three amobarbital injections (120 mg/kg) while the animals were maintained on ad libitum food and water. When tested while food deprived, so that any previously conditioned drive cues were absent, these animals still showed an attenuated aversion relative to no-preexposure controls. This experiment shows that conditioning an aversion to a drive state (Peck & Ader, 1974) is not a necessary condition for obtaining the UCS preexposure effect (cf. Braveman, 1975; Elkins, 1974).

Several studies using the conditioned taste aversion (CTA) paradigm not only have obtained a preexposure effect but also have found no evidence of conditioning even with repeated taste–drug pairings (Berman & Cannon, 1974; Cappell & LeBlanc, 1977; LeBlanc & Cappell, 1974). Brookshire and Brackbill (1976), for example, preexposed one group of rats to apomorphine injections (15 mg/kg) for 10 consecutive days, whereas a second group of rats received saline injections. The group preexposed to apomorphine formed no aversion to saccharin paired with an injection of apomorphine (15 mg/kg), even after repeated taste–drug pairings. Similarly, Cappell and LeBlanc (1975) showed that rats that had received 20 prior exposures to D-amphetamine (7.5 mg/kg) subsequently failed to develop an aversion to saccharin that was repeatedly paired with D-amphetamine (1.0 mg/kg). On the other hand, one or five prior exposures to the drug retarded, but did not prevent, the acquisition of a taste aversion (cf. Goudie et al., 1976). These findings contrast with experiments on human and rabbit eyelid conditioning and the CER, in which repeated CS–UCS pairings administered after preexposure to the UCS eventually yield a level of conditioning approxi-

mately equal to that of no-preexposure controls (Hobson, 1968; Kamin, 1961; Mis & Moore, 1973).

Riley, Jacobs, and LoLordo (1976) suggested that failures to obtain any conditioning with repeated taste–drug pairings following preexposure to the UCS may be restricted to addictive drugs (e.g., apomorphine, ethanol, and morphine) or to those drugs that animals will self-administer (e.g., amphetamine). They demonstrated that preexposure to six injections of LiCl (.15 M) attenuated an aversion established by a few pairings of saccharin and LiCl (.15 M), relative to no-preexposure controls. However, repeated saccharin–LiCl pairings eventually produced a substantial aversion to saccharin even in rats preexposed to LiCl. In a similar experiment, Cannon et al. (1975) preexposed groups of rats to a single LiCl intubation (.12 M) 1, 4, or 8 days prior to seven conditioning trials in which saccharin was paired with LiCl (.12 M). A single dose of LiCl given 1 day, but not 4 or 8 days, prior to conditioning attenuated the aversion to saccharin established by a single taste–drug pairing (cf. Cappell & LeBlanc, 1975, 1977). However, repeated taste–drug pairings induced a substantial aversion to saccharin even in the group given exposure to LiCl 1 day prior to conditioning.

Holman (1976) preexposed groups of rats to eight injections of either LiCl (.15 M) or saline and then conditioned an aversion by pairing the taste of sodium chloride (NaCl) with LiCl (.15 M). The group preexposed to LiCl showed a retarded acquisition of the aversion to NaCl, but substantial conditioning occurred with repeated taste–drug pairings.

Thus, the results of experiments with LiCl, a nonaddictive drug that animals will not self-administer, lend some support to the contention of Riley et al. that failures to obtain conditioning with repeated pairings of tastes and amphetamine, morphine, apomorphine, or ethanol following preexposure to these drugs reflect their addictive qualities, or animals' tendency to self-administer them. It is unfortunate, however, that none of the above experiments performed a direct comparison between addictive and nonaddictive drugs.

Several experiments have evaluated the effect of preexposure to signaled versus unsignaled drug UCSs on the subsequent formation of a taste aversion based on these drug UCSs. Cannon et al. (1975) showed that a single, signaled preexposure to LiCl does retard the acquisition of an aversion, although the decremental effect is smaller than that produced by a single, unsignaled preexposure to LiCl. In this experiment, one group of rats received a single sustacal–LiCl (.12 M) pairing, a second group of rats received a water–LiCl (.12 M) pairing, and a third group of rats received a sustacal–saline pairing. All groups then received a saccharin–LiCl (.12 M) pairing on the following day and on 3 additional days. Signaled and unsignaled preexposure to LiCl equally attenuated the aversion induced by saccharin–LiCl pairings on the first conditioning trial. However, repeated saccharin–LiCl pairings produced a stronger aversion in the signaled preexposure group than in the unsignaled preexposure group.

Mikulka, Leard, and Klein (1977) found an even more striking difference between the effects of signaled and unsignaled preexposure conditions. They observed that unsignaled preexposure to LiCl retarded subsequent conditioning even when signaled preexposure had no such effect. Saline-preexposed controls and a group that received four prior pairings of sucrose and LiCl acquired an aversion to almond-flavored water paired with LiCl equally rapidly, whereas a group that received four unsignaled exposures to LiCl showed a somewhat attenuated aversion to almond-flavored water.

Not all studies have obtained differences between signaled and unsignaled preexposure conditions. Goudie, Thornton, and Marsh (Note 3) found that unsignaled preexposure to methamphetamine and preexposures signaled by access to a saline solution had equal attenuating effects on subsequent conditioning of an aversion to saccharin paired with methamphetamine. Zellner and Riley (Note 4) obtained similar results with the stimulant drug methylphenidate. Goudie et al.'s failure

to obtain a difference between signaled and unsignaled preexposure conditions led them to consider the importance of the relationship between responding and the UCS during the preexposure phase. They suggested that during the preexposure phase, rats learn that the occurrence of the UCS is independent of their behavior. This learning, called "learned helplessness" by Maier and Seligman (1976), interferes with subsequent learning to withhold drinking in the presence of a taste cue that reliably precedes some UCS. Goudie et al. tested this notion by allowing groups of rats ad libitum access to only water or only methamphetamine during the preexposure phase. When rats were given some control over the preexposed UCS, that is, when the presentation of the UCS depended on responding, the attenuating effect of preexposure to the UCS was abolished (cf. Deutsch & Eisner, 1977). Recently, however, Randich (1978) reported that rats permitted to escape electric shock presentation during a preexposure phase and thereby to control the UCS showed attenuation of subsequent CER conditioning, even greater attenuation than yoked controls that could not escape electric shock during the preexposure phase.

Rudy, Iwens, and Best (1977) performed a series of experiments that showed that pairing a novel, exteroceptive background stimulus with poison during preexposure attenuates the development of a CTA regardless of whether the background stimulus is present or absent when excitatory conditioning is performed.

In Experiment 1, three groups of rats were preexposed to injections of LiCl (.15 M) for 4 days. Two of these groups received LiCl injections in the presence of a novel exteroceptive stimulus (a black chamber). The rat was first placed in the black chamber for 5 minutes, was then removed and injected with LiCl, and was finally replaced in the black chamber for an additional 25 minutes. A third group of rats received LiCl injections in their home cages. A fourth group of rats were simply placed in the black chamber and not injected with LiCl. Following the preexposure phase, one of the groups preexposed to poison in the black chamber received a single pairing of a saccharin infusion and a LiCl injection in the home cage. The other three groups received a single pairing of a saccharin infusion and a LiCl injection in the black chamber and were then removed to the home cage. A two-bottle test between saccharin and water in the home cage revealed that animals preexposed to LiCl in the home cage and conditioned in the black chamber did not significantly differ from the group that received no preexposure treatment; that is, the normal UCS preexposure effect was not obtained. However, preexposure to poison in the black chamber did attenuate the aversion to saccharin whether the saccharin–LiCl pairing occurred in the home cage or in the black chamber.

In a second experiment, Rudy, Iwens, and Best ruled out the possibility that the group preexposed to LiCl in the home cage and conditioned in the black chamber failed to show a preexposure effect because the novel stimulation produced by placement in the black chamber during conditioning in some way potentiated the effect of LiCl.

A third experiment showed that the 5-minute preinjection placement, rather than the 25-minute postinjection placement, in the black chamber during preexposure was responsible for the attenuation effect obtained in Experiments 1 and 2.

In a final experiment, two groups of rats received the preexposure to LiCl in the presence of either the novel black chamber or a novel light stimulus. A third group of rats received the preexposure treatment with LiCl in the presence of the black chamber after being familiarized with the black chamber by four prior placements in it. A fourth group of rats received LiCl injections in the home cage. All groups then received a single saccharin infusion paired with LiCl in the home cage. The results indicated that only the groups preexposed to LiCl in the presence of the unfamiliar, novel stimuli, that is, the black chamber or the light stimulus, showed the attenuation effect.

Rudy, Iwens, and Best found that rats preexposed to LiCl in the presence of a novel exteroceptive stimulus subsequently displayed

an attenuated taste aversion relative either to rats that received no prexposure treatment or to rats preexposed to LiCl in a familiar environment. They concluded that this attenuation effect was not a result of blocking by prior conditioning of the novel exteroceptive stimulus during preexposure to LiCl, because the attenuation occurred in groups preexposed to LiCl in the presence of a novel stimulus whether that stimulus was present or absent during subsequent conditioning of the aversion. If the novel exteroceptive stimulation was not present during saccharin–LiCl pairings, it could not have blocked conditioning to saccharin.

Instead, Rudy, Iwens, and Best (1977) argued that the handling and infusion procedure acquired associative strength during preexposure to LiCl and attenuated subsequent taste aversion learning. This explanation demands that the groups preexposed to LiCl in the presence of the novel exteroceptive stimulus more strongly associated the handling cues with poison than the groups preexposed to LiCl in the home cages, which were handled in the same way. For this to occur, Rudy, Iwens, and Best first assumed that handling cues were latently inhibited for all groups during the initial stages of the experiment, that repeated occurrence of the handling and infusion procedure prior to the preexposure phase rendered these cues less associable with UCSs. Placement in the novel chamber disrupted the latent inhibition such that the handling cues that would not otherwise have been associated with poison did acquire the associative strength necessary to attenuate the formation of a taste aversion.

In another series of experiments, Rudy, Rosenberg, and Sandell (1977) obtained evidence that supports this sort of account. Presentation of a novel exteroceptive cue just prior to the pairing of a familiar taste with LiCl injection enhanced the conditioning of aversion to the taste, an outcome formally similar to that postulated by Rudy, Iwens, and Best for familiar handling (rather than taste) cues.

Cross-UCS Comparisons

Gamzu (Note 5) preexposed two groups of rats to either D-amphetamine (2 mg/kg) or chlordiazepoxide (15 mg/kg) and two groups of rats to saline. The two drug-preexposed groups and one saline-preexposed group then received pairings of saccharin and D-amphetamine, while the other saline-preexposed group received pairings of saccharin and saline. The group preexposed to D-amphetamine showed an attenuated aversion to saccharin when compared with the group preexposed to saline and conditioned with D-amphetamine. However, the group preexposed to chlordiazepoxide and conditioned with D-amphetamine showed an aversion comparable with that of the no-preexposure controls. Thus, preexposure to D-amphetamine attenuates a conditioned aversion when this drug is used as the UCS, but preexposure to chlordiazepoxide does not affect the aversion induced by D-amphetamine.

Goudie and Thornton (1975) preexposed two groups of rats to either D-amphetamine (2 mg/kg) or *dl*-fenfluramine (6 mg/kg) for 9 consecutive days. Five days elapsed between the last UCS preexposure and the start of excitatory conditioning. During conditioning, half the rats in each group received pairings of saccharin and D-amphetamine, and the other half received pairings of saccharin and *dl*-fenfluramine. Preexposure to D-amphetamine attenuated the aversion induced by pairings of saccharin and D-amphetamine relative to the appropriate controls, but did not attenuate the aversion induced by pairings of saccharin and *dl*-fenfluramine. However, preexposure to *dl*-fenfluramine attenuated the aversion induced by either pairings of saccharin and *dl*-fenfluramine or pairings of saccharin and D-amphetamine.

Vogel (Note 2) preexposed groups of rats to either D-amphetamine (2 mg/kg), amobarbital (120 mg/kg), or the vehicle for 3 days. Then all groups were given taste–drug pairings in a factorial design. Preliminary work indicated that these dosages of D-amphetamine and amobarbital produced comparable aversions. Vogel found that preexposure to D-amphetamine did not attenuate

the aversion induced by pairing a taste and D-amphetamine, a result that runs counter to the results of Gamzu (Note 5) and Goudie and Thornton (1975). However, preexposure to D-amphetamine did attenuate the aversion induced by taste–amobarbital pairings. Furthermore, preexposure to amobarbital attenuated the aversion induced by taste–amobarbital pairings, but did not attenuate the aversion induced by taste–D-amphetamine pairings.

Cappell, LeBlanc, and Herling (1975) showed that chronic preexposure to D-amphetamine (20 mg/kg) administered over a 20-day period attenuated the aversion induced by pairing saccharin and morphine (6 mg/kg). Further, there was no evidence of conditioning even with repeated taste–drug pairings. However, preexposure to morphine facilitated an aversion induced by pairing saccharin and D-amphetamine (1 mg/kg). In a third experiment, chronic preexposure to chlordiazepoxide (25 mg/kg) for 22 days had no effect on the aversion induced either by pairing saccharin and D-amphetamine (1 mg/kg), a result consistent with that reported by Gamzu (Note 5), or by pairing saccharin and morphine (6 mg/kg). However, pretreatment with chlordiazepoxide did attenuate an aversion induced by pairing saccharin and chlordiazepoxide (cf. Gamzu, 1977).

Whaley, Scarborough, and Reichard (1966) preexposed two groups of rats to 1,000 rotations in a tumbling apparatus, whereas two other groups of rats were simply placed in the tumbling apparatus without being rotated. Twenty-four hours later, half the rats in each of these groups were irradiated with X rays (6 r./min.) for 10 minutes, while the other half in each group were sham irradiated. Immediately following this treatment, all rats received free access to tap water and saccharin for 20 minutes, so that saccharin consumption preceded some of the delayed effects of x-irradiation for rats in the experimental groups. A subsequent test found that the experimental group that had been tumbled showed an attenuated saccharin aversion relative to the irradiated group that had not been tumbled.

Braveman (1975) examined cross-UCS effects by preexposing groups of rats to injections of scopolamine methyl nitrate (1 mg/kg) for 0, 1, 3, 5, or 7 days and then giving them a single pairing of saccharin and LiCl (.3 M). Groups that were preexposed to scopolamine five or seven times consumed reliably more saccharin than the no-preexposure control group in a test following the saccharin–LiCl pairing. The amount of saccharin consumed in the test was a direct function of the number of preexposures to scopolamine. In a later study, Braveman (1977) also observed that preexposure to LiCl attenuated the magnitude of an aversion to a taste paired with scopolamine methyl nitrate.

Braveman (1975) considered the possibility that these drugs acted on the same physiological substrate and did not truly represent a cross-UCS effect. He attempted to eliminate this possibility by preexposing rats to either D-amphetamine (2 mg/kg) or scopolamine methyl nitrate (1 mg/kg), two drugs that induce conditioned taste aversions via different physiological substrates according to the ablation data of Berger, Wise, and Stein (1973). Rats that were preexposed to D-amphetamine then received a pairing of saccharin and scopolamine methyl nitrate, whereas rats that were preexposed to scopolamine methyl nitrate received a pairing of saccharin and D-amphetamine. In general, groups that were preexposed to a drug showed attenuated saccharin aversions relative to nonpreexposed controls. On the basis of this cross-UCS effect, Braveman concluded that drug preexposure did not modify the physiological mechanism that underlies conditioned food aversions.

Braveman also considered the possibility that cross-drug tolerance effects produced these results. He sought to eliminate this possibility by using mechanical rotation, rather than a drug, as a UCS. Groups of rats were either preexposed to rotation at 60 rpm for 15 minutes/day or injected with (a) scopolamine methyl nitrate, (b) D-amphetamine, or (c) LiCl for 5 days. All groups then received repeated pairings of saccharin and mechanical rotation. All of the groups preexposed to a UCS failed to show any aversion to saccharin, that is, they drank as much as controls that received neither preexposure nor

conditioning with the UCS. In a later study using a factorial design, Braveman (1977) preexposed rats to either LiCl injections or a severe electric shock treatment and a week later presented them with a taste followed by either LiCl or electric shock. A symmetrical cross-UCS preexposure effect was obtained. Braveman concluded that preexposure to any aversive UCS results in diminution of a general stress response to aversive stimuli, thereby reducing their impact during subsequent conditioning. Braveman's hypothesis is very similar to the emotional reactivity explanation of the preexposure effect suggested by other investigators (Kamin, 1961; Mis & Moore, 1973; Taylor, 1956).

Cannon et al. (1977) repeatedly preexposed rats to LiCl, ethanol, or NaCl and then presented half the rats in each group with pairings of saccharin and LiCl, whereas the others received saccharin followed by ethanol. Relative to saline-preexposed controls, rats that had been preexposed to one drug and then conditioned with another showed a strong UCS preexposure effect. However, the cross-UCS preexposure effect was significantly smaller than the within-UCS effect, leading Cannon et al. to conclude that in addition to some general mechanism that mediates cross-UCS preexposure effects, drug-specific components also contribute to the effect when the same drug is used during both preexposure and conditioning. As noted earlier, Cannon et al. suggested that tolerance contributed to the drug-specific effect, at least in the case of ethanol.

Summary of Studies Assessing Preconditioning Exposure to a UCS

The results of studies assessing the effect of preconditioning exposure to the UCS on human and rabbit eyelid conditioning, CER conditioning, and conditioning of taste aversions are briefly summarized below. The decremental effect of preconditioning exposure to the UCS on the formation of an excitatory CR is

1. In the human eyeblink, rabbit nictitating membrane, and taste aversion procedures a direct function of the intensity or concentration of an unsignaled preexposed UCS (Cannon et al., 1975; Mis & Moore, 1973; Taylor, 1956) and an inverse function of the intensity or concentration of the UCS used in excitatory conditioning (Cannon et al., 1975).
2. In the CER procedure, an inverted-U-shaped function of the intensity of an unsignaled (or signaled) preexposed UCS (Randich, 1978).
3. A direct function of the number of preexposed UCSs (Cannon et al., 1975; Elkins, 1974; Goudie et al., 1976; Hobson, 1968; Mis & Moore, 1973; Randich, 1978; Vogel, Note 2), even when cross-UCS comparisons are made (Braverman, 1975).
4. An inverse function of the time interval between the last preexposure to a UCS and the start of excitatory conditioning (Cannon et al., 1975; Cappell & LeBlanc, 1975, 1977; Mikulka et al., 1977; Mis & Moore, 1973).
5. Sometimes reduced if the preexposed UCS is signaled (Cannon et al., 1975; Mikulka et al., 1977; see also, Revusky, Parker, Coombes, & Coombes, 1976), although Goudie et al. (Note 3) and Zellner and Riley (Note 4) found no differences between signaled and unsignaled preexposure to the UCS on subsequent taste aversion learning (cf. Randich, 1978).
6. Not overcome by repeated CS–UCS pairings when some addictive drugs and amphetamine are used (Berman & Cannon, 1974; Brookshire & Brackbill, 1976; Cappell & LeBlanc, 1977; Cappell et al., 1975; LeBlanc & Cappell, 1974), but is overcome when nonaddictive drugs are used (Cannon et al., 1975; Holman, 1976; Riley et al., 1976).
7. Often independent of the type of aversive UCS used in the preexposure and conditioning phases (Braveman, 1975, 1977; Cappell et al., 1975; Goudie & Thornton, 1975; Whaley et al., 1966; Vogel, Note 2; Gamzu, Note 5).

Theoretical Considerations

In general, the studies examined in this article show that preconditioning exposure to aversive UCSs retards the formation of an excitatory CR. Further, this effect often occurs when the UCS used during the pre-

exposure phase is different from the UCS used during excitatory conditioning (cf. Braveman, 1975). The basis of this phenomenon remains an enigma, although both associative and nonassociative theories have been advanced. Many of these theories are of limited scope and confine their predictive value to particular situations, for example, those in which the UCS is an addictive drug (cf. Parker et al., 1973). This is not to say that such theories are incorrect; rather, the generality of the preexposure phenomenon across Pavlovian conditioning paradigms, species, and aversive UCSs invites a more general interpretation. It is toward this end that theories of the UCS preexposure effect are evaluated in the following sections.

Associative Theories

Blocking. As was noted earlier, a blocking interpretation of the UCS preexposure phenomenon posits that some stimulus aspect (X) of the experimental situation acquires associative strength during preexposure to an aversive UCS. Conditioning of Stimulus X reduces the amount of associative strength that a nominal stimulus (A) can acquire during subsequent excitatory conditioning in that same environment (AX). As a result, the rate of acquisition of an excitatory CR to Stimulus A is attenuated.

There are three critical assumptions of a blocking interpretation of the UCS preexposure effect. First, Stimulus X can be any aspect of the experimental situation. For instance, Stimulus X may be static cues provided by the characteristics of the experimental environment, cues provided by the handling procedure, or cues provided by the injection procedure in CTA experiments. Second, Stimulus X must be present during both the preexposure and the excitatory conditioning phases of an experiment for blocking to occur. Third, the presence of the previously conditioned Stimulus X interferes with conditioning of Stimulus A because the UCS can support only a limited amount of associative strength (Rescorla & Wagner, 1972) or because there is little conditioning to redundant predictors of reinforcement (Mackintosh, 1975). There are several predictions that can be derived from a blocking account of the UCS preexposure phenomenon:

1. Blocking predicts retarded acquisition of an excitatory CR to Stimulus A if conditioning of Stimulus X occurs during preexposure to the UCS and Stimulus X is present during excitatory conditioning of Stimulus A. Virtually all the studies evaluated in this article obtained an attenuation of excitatory conditioning following prior exposure to the UCS. These studies typically included some stimulus (X) during both the preexposure and the conditioning phases. This stimulus could block conditioning of Stimulus A. The questions at hand are whether such stimuli (X) are in fact conditionable and whether such conditioning is capable of blocking conditioning of a discrete Stimulus A.

Many of the aforementioned stimuli that can be identified as Stimulus X are conditionable. Tomie (1976) demonstrated that prior presentations of free food attenuate the rate of acquisition of an autoshaped keypeck response. Tomie's data suggest that conditioning of contextual stimuli during free-food presentations blocks conditioning of a nominal CS subsequently presented in the same context. Similarly, Willner (1978) demonstrated that rats preexposed to injections of LiCl in a distinctive environment develop a place aversion and that this conditioning attenuates the formation of an aversion to saccharin that is later paired with LiCl in the same place. In two separate experiments using D-amphetamine and LiCl as UCSs, Braveman (Note 6) preexposed and conditioned rats in either the same or different environments; the environments were distinguished by the level of auditory–visual stimulation. Braveman obtained a UCS preexposure effect only when preexposure and conditioning occurred in the same environment, supporting an associative explanation of the phenomenon.

However, it is not clear whether conditioning of contextual cues occurs in other preparations, such as in human and rabbit eyelid conditioning. Mis and Moore (1973) argued against a context-blocking interpre-

tation of their rabbit eyelid conditioning study, citing the lack of evidence that the nictitating membrane response in rabbits can be conditioned to contextual stimuli. Their claim may not bear directly on a general blocking view of the UCS preexposure phenomenon. Preexposure to a UCS may neither alter the specific UCR that an experimenter measures nor condition that UCR to contextual cues, but may instead condition a fear response elicited by the UCS to contextual cues. If this fear response is important for excitatory conditioning of the nictitating membrane response (cf. Konorski, 1967), then a blocking interpretation may still be useful. In fact, Hinson and Siegel (Note 7) have recently presented evidence in support of a context-blocking interpretation of the UCS preexposure effect in nictitating membrane conditioning.

Researchers have attempted to reduce the likelihood of context blocking by providing a nominal signal (CS) for the UCS during the preexposure phase. Insofar as conditioning occurs to the salient, nominal CS it will not occur to less salient contextual cues (cf. Kamin, 1969). Hence, no blocking should occur when a novel CS is subsequently paired with the UCS in the same context. Cannon et al. (1975) and Mikulka et al. (1977) showed that signaled preexposure either attenuates or eliminates completely the retardation of excitatory conditioning that is typically obtained with unsignaled preexposure conditions. These results are compatible with a blocking interpretation because contextual stimuli should acquire little or no associative strength during preexposure to signaled UCSs. However, Goudie et al. (Note 3) and Zellner and Riley (Note 4) reported that rats given signaled or unsignaled preexposure to stimulants showed equal attenuation of the acquisition of a taste aversion, relative to controls. Randich (1978) also reported that rats given signaled preexposure to an electric shock UCS showed greatly retarded acquisition of a CER, relative to appropriate controls.

A second source of stimulation that could potentially block conditioning of a nominal CS is the handling procedure. Rudy, Iwens, and Best (1977) specifically attempted to condition contextual cues by preexposing their rats to illness in the presence of a novel, exteroceptive stimulus, namely, a black chamber. This preexposure treatment attenuated the formation of a CTA to a greater extent than did preexposure to illness in the home cage. However, the effect of preexposure to the UCS in a novel environment occurred regardless of whether the novel exteroceptive stimulus was present or absent when taste–drug pairings were administered. This result seems incompatible with a blocking interpretation, which asserts that Stimulus X must be present during excitatory conditioning to block acquisition of conditioning to Stimulus A. Rudy, Iwens, and Best performed subsequent experiments that forced them to conclude that blocking was indeed responsible for their results, although not as a result of associative strength acquired by the novel, exteroceptive stimulus. They argued that the novel exteroceptive stimulus acted only as a disinhibitor of other latently inhibited background cues (the handling cues in particular), thus allowing these cues to acquire associative strength. Since the handling cues were present during pairings of Stimulus A with the UCS, they blocked conditioning of an aversion to Stimulus A.

Braveman (1978) reasoned that if conditioning to handling cues during drug preexposure were responsible for the UCS preexposure effect, then manipulations that should attenuate conditioning to handling cues should also attenuate the UCS preexposure effect. He varied the amount of handling received by groups of rats prior to preexposure to LiCl and found no evidence that this variable influenced the UCS preexposure effect, which was strong even in rats that had been handled for 21 days prior to preexposure. However, Braveman did not administer any saline injections prior to drug preexposure, leaving open the possibility that conditioning to injection cues was responsible for the UCS preexposure effect. Recent evidence (Willner, 1978; Poulos & Cappell, Note 8) indicates that cues associated with the injection of a drug can acquire associative strength and are capable of blocking

conditioning of an aversion to a distinctive taste. This conclusion is based on the finding that the normal UCS preexposure effect obtained by prior injections of a drug can be reduced by degrading the correlation between injection cues and the drug during the preexposure phase.

2. Blocking predicts that the effect of preexposure to the UCS should dissipate with the time to the start of excitatory conditioning, as long as the animal is maintained in the presence of Stimulus X during this period. This prediction is based on the view that the conditioned response to Stimulus X will extinguish during a delay period in which it is not reinforced. If the CR to Stimulus X is near zero at the start of excitatory conditioning, then blocking should not occur.

Mis and Moore (1973), Cappell and LeBlanc (1975, 1977), and Cannon et al. (1975) have shown that the magnitude of the UCS preexposure effect decreases as the time interval increases between the last preexposure to a UCS and the start of excitatory conditioning. In the CTA studies, the animals spent the delay period in the preexposure environment, that is, in the presence of Stimulus X. It is unclear whether this outcome reflects extinction of Stimulus X or some effect that is independent of context, since the experiments included no conditions in which the animals spent the delay interval in the absence of Stimulus X. A critical test of this prediction would involve removing half the preexposed animals from the preexposure environment. The group removed from the preexposure environment during the delay period should show a greater attenuation of excitatory conditioning than the group maintained in the preexposure environment.

In a procedure that is formally similar to that just described, Batson and Best (Note 9) preexposed rats to injections of LiCl after they were placed in a distinctive black box. Then over the next 8 days, half the preexposed rats received 16 trials in which they were placed in the black box and then injected with physiological saline. These trials should have extinguished any conditioned response to the black box and associated cues. The other rats remained in their home cages during this period. Then both groups received a single saccharin–LiCl pairing following placement in the black box. Finally, both groups and the nonpreexposed controls were permitted to drink saccharin in the home cage. Rats that had received extinction trials were as averse to saccharin as were nonpreexposed controls, but rats that had spent the interval between preexposure and conditioning in the home cage showed a strong UCS preexposure effect. Batson and Best concluded that associative blocking formed the basis of their UCS preexposure effect (cf. Hinson & Siegel, Note 7, for a similar manipulation and outcome).

3. Blocking predicts that an animal preexposed to the UCS should eventually attain the same level of excitatory conditioning as an animal not preexposed to the UCS, although this level of conditioning should be attained at a slower rate. This should occur because the UCS is presented only in the presence of the nominal CS during excitatory conditioning and never in its absence. According to the Rescorla–Wagner (1972) model, this should result in more nonreinforced than reinforced presentations of Stimulus X, and the associative strength of Stimulus X should decline to zero. The nominal Stimulus A will acquire as much associative strength as Stimulus X loses. Thus, for an animal preexposed to the UCS, Stimulus A is permitted to acquire as much associative strength as the Stimulus A for an animal not preexposed to the UCS. Mackintosh's (1975) model makes the same prediction, asserting that Stimulus X should lose associative strength because it is a poorer predictor of the UCS than is Stimulus A. Thus, the salience of Stimulus A will increase, and Stimulus A will gain associative strength. Preexposed and control groups have attained the same level of conditioning in several studies that continued to present CS–UCS pairings during excitatory conditioning (Holman, 1976; Kremer, 1971; Mis & Moore, 1973). However, a few studies failed to obtain any conditioning following preexposure to a drug, even with repeated CS–UCS pairings (Berman & Cannon, 1974; Braveman, 1975; Brookshire & Brackbill, 1976; Cap-

pell & LeBlanc, 1977; Cappell et al., 1975). These are notable exceptions, however, because all but Braveman's study involved the use of UCSs that an animal will self-administer, namely, D-amphetamine, morphine, and ethanol.

UCS controllability. Vogel (Note 2) and Goudie et al. (Note 3) proposed that an organism that is preexposed to an aversive UCS learns that the aversive state is uncontrollable. This learning produces an associative deficit that transfers to the excitatory conditioning phase and interferes with the learning of the relationship between the CS and the UCS. This hypothesis is an extension of the learned helplessness notion (Maier & Seligman, 1976), which essentially argues that the organism perceives the occurrence of the UCS during the preexposure phase as uncorrelated with anything it does or attempts to do. Thus, it is possible that the associative deficit produced by learning that either stimuli or responses are ineffective in predicting or controlling the UCS interferes with the formation of an association between the stimulus and the reinforcer during excitatory conditioning.

Data showing that signaled preexposure to a UCS reduces the preexposure effect (Cannon et al., 1975) could be interpreted to support this contention. This, of course, assumes that an organism can learn some form of preparatory response during signaled preexposure to a UCS that minimizes and thereby controls the impact of the UCS.

On the other hand, Goudie et al. (Note 3) found equal attenuating effects of signaled and unsignaled preexposure to methamphetamine on the formation of a subsequent taste aversion. They suggested that since the UCS is uncontrollable in both unsignaled and signaled preexposure conditions, controllability of the UCS may be important. Subsequently these authors demonstrated that rats allowed ad libitum access to methamphetamine during the preexposure phase do not show attenuated saccharin aversions when saccharin is paired with methamphetamine injections. These data are compatible with the hypothesis that controllability of UCS onset is the important determinant of the UCS preexposure effect. A test of this hypothesis can only be accomplished by using drugs that an animal will self-administer, for example, methamphetamine. The results of such a test may be applicable only to experiments that use drugs of abuse. On the other hand, a related hypothesis that focuses on the controllability of UCS termination can be tested with a variety of aversive UCSs. In this regard, Randich (1978) found that rats permitted to terminate electric shocks during a preexposure phase showed even greater attenuation of subsequent CER conditioning than yoked controls not permitted to terminate electric shocks during the preexposure phase. Thus, giving an animal control over the termination of an aversive UCS does not eliminate the UCS preexposure phenomenon.

The UCS onset-controllability hypothesis would account for symmetrical cross-UCS effects (cf. Goudie & Thornton, 1975) if one assumed that the aversive states induced by drugs were similar in being equally uncontrollable. Of course, given this assumption, asymmetrical cross-UCS effects pose a problem, although they primarily involve the use of D-amphetamine, a drug that has both positive and negative reinforcing characteristics (Wise, Yokel, & DeWit, 1976).

Nonassociative Theories

Artificial need states. Parker et al. (1973) suggested that preexposure to addictive drugs may induce an artificial need state for these drugs. The failure to obtain a CTA would then reflect the fact that the CS predicts alleviation of withdrawal symptoms correlated with the need state. This hypothesis is of limited general applicability. In addition, LeBlanc and Cappell (1974) have shown that preexposure to amphetamine attenuates the formation of a CTA but does not induce a need state for amphetamine.

UCS novelty. Amit and Baum (1970) and Gamzu (1977; Gamzu, Note 5) suggested that the novelty of the UCS is an important determinant of associative learning. Any treatment, such as preexposure to the UCS, that reduces the novelty of the UCS

should retard the acquisition of an excitatory CR. The usefulness of this hypothesis is directly questioned by both symmetrical and asymmetrical cross-UCS effects. If one assumes that this hypothesis can be restated as "the novelty of the aversive state is important for associative learning," then it may be possible to account for symmetrical cross-UCS effects in terms of a common aversive reaction to the UCSs. However, in doing so it becomes difficult to account for asymmetrical cross-UCS effects. In this regard, Gamzu (1977) argued that asymmetrical cross-UCS effects may reflect quantitative rather than qualitative effects of the two drugs involved. For example, he argued that the dose of *dl*-fenfluramine used by Goudie and Thornton (1975) to attenuate conditioning of a taste aversion to *dl*-fenfluramine was more aversive than the dose of D-amphetamine (see also, Braveman, 1977; Cannon et al., 1977).

Although such an argument has some force, the main problem with the UCS-novelty account is that it has not been specified in enough detail to be easily tested.

Central habituation. Perhaps the most widely accepted nonassociative explanation of the UCS preexposure phenomenon is that some central habituation process occurs in response to repeated applications of the UCS during the preexposure phase and reduces the organism's responsiveness to subsequent applications of the UCS (Kamin, 1961; Mis & Moore, 1973; Taylor, 1956). This hypothesis has been stated in a variety of forms. For example, Braveman (1975) stated that an organism may habituate to the stress induced by the UCS. Similarly, it has been suggested that an organism may develop a tolerance for the UCS when drugs are used as UCSs (Cappell et al., 1975; LeBlanc & Cappell, 1974; Riley et al., 1976). Central habituation to the UCS, habituation to the stress induced by the UCS, and tolerance for the UCS are formally quite similar accounts.

It is possible that some physiological substrate, which is activated by an aversive UCS and is important for excitatory conditioning, is modified by repeated exposure to the UCS. Riley et al. (1976) suggested that this physiological substrate may involve the pituitary–adrenal axis and, in particular, the adrenocorticotrophic hormone (ACTH). For example, ACTH has been shown to be critical for normal acquisition of conditioned active and passive avoidance responses (cf. DeWied, 1964), and the release of ACTH is conditioned in CER paradigms (Bassett, Cairncross, & King, 1973; Brady, 1967) and CTA paradigms (Ader, 1977). It is also known that exposure to an unsignaled, aversive UCS often inhibits the normal release of ACTH from the adenohypophysis in response to subsequent applications of that UCS or to other aversive stimuli (Milulaj & Mitro, 1972; Munson, 1973). It is possible, therefore, that inhibition of stress-induced release of ACTH by prior exposure to unsignaled, aversive stimuli would reduce the potentiating effect that ACTH has on the acquisition and maintenance of fear-motivated behaviors. The possibility that ACTH plays a role in the preexposure phenomenon is intriguing for the following reasons. First, it would provide a common final substrate through which many aversive stimuli might act and thus a basis for the notion of central habituation. Second, some investigators (Rescorla, 1973) have obtained evidence of a preexposure effect only during extinction testing, a situation in which Weiss, McEwen, Silva, and Kalkut (1970) have shown ACTH to have maximal effects on behavior. However, the hypothesis that ACTH plays a critical role in the UCS preexposure phenomenon should be viewed with some caution, considering the lack of data.

Memorial representation. Rescorla (1974) proposed that first-order excitatory conditioning involves the construction of memories for individual events and a formation of associations (stimulus–stimulus) between such memories. Preconditioning exposure to a UCS produces a memorial representation of that UCS that may either augment or diminish the representation of the UCS used during excitatory conditioning. An unsupplemented memorial-representation model predicts no UCS preexposure effect unless UCS intensity is changed between the preexposure and conditioning phases and is thus untenable. Consequently, it seems worthwhile to

consider how central habituation and changes in memorial representation of the UCS might interact. It may be the case, for example, that animals preexposed to any intensity UCS will show an attenuation of excitatory conditioning because of central habituation but that the amount of attenuation will also be a function of changes in the memorial representation of the UCS. In general, Rescorla's view would predict that the attenuating effect of UCS preexposure should be diminished when the UCS is more intense during preexposure than during excitatory conditioning, but enhanced when the UCS is less intense during preexposure than during conditioning. In studies that used unsignaled preexposure conditions (Cannon et al., 1975; Mis & Moore, 1973) and signaled preexposure conditions (Randich, 1978), these results were not obtained. Thus, changes in the memorial representation of the UCS are unlikely to contribute to the UCS preexposure effect.

Opponent-process theory. Another nonassociative account of the decremental effect of preexposure to the UCS comes from the opponent-process theory of acquired motivation (Solomon, 1977; Solomon & Corbit, 1974)—a theory designed to explain the affective dynamics of responses to strong stimuli. The opponent-process theory holds that the relative strengths of two opposing processes, the a process and the b process, determine the affective state of an organism in response to a strong stimulus. The a process, or primary affective process, is postulated to be the emotional UCR to the UCS, for example, fear elicited by a strong electric shock. The a process is said to be stimulus locked, showing little habituation or sensitization with successive presentations of the UCS. The b process, or the opponent process, is said to be aroused by occurrence of the a process and to have an affective sign opposite to that of the a process, for example, inhibition of fear following strong shock. The b process, unlike the a process, is postulated to increase in both intensity and duration with repeated evocation, as long as the time interval between successive evocations is less than that required for the complete decay of the b process (*critical decay duration;* Starr, 1976). The algebraic summation of these two opposing affective processes results in a standard pattern of affective dynamics. When the quantity $a - b$ is positive, the animal is said to be in the A state; when the quantity $a - b$ is negative, the animal is said to be in the B state.

This model bears on the UCS preexposure phenomenon because it predicts that repeated presentations of the UCS will strengthen the b process when these presentations are separated by less than the critical decay duration. As this occurs algebraic summation of the opponent processes should result in a greatly diminished A state and in an augmented B state with a long decay time. If excitatory classical conditioning is the conditioning an A state to a nominal CS, then such conditioning will be attenuated by prior exposure to the UCS because this treatment diminishes the A state.

The opponent-process model makes a prediction that is not made by other nonassociative accounts of the UCS preexposure phenomenon. If the b process grows with repeated UCS presentations, then the conditioning of the B state should be facilitated by prior exposure to the UCS. Thus, if a nominal CS is paired with the peak of the B state in a backward conditioning procedure, conditioning of this B state should be facilitated by prior exposure to the UCS. Other nonassociative accounts of the UCS preexposure phenomenon that have been discussed predict attenuating effects of UCS preexposure on both excitatory (A state) and inhibitory (B state) conditioning. These differential predictions have not yet been tested.

Summary

The first researchers to demonstrate the decremental effect of prior exposure to the UCS on the acquisition of an excitatory CR explained this decrement in nonassociative terms (Kamin, 1961; Kimble & Dufort, 1956; Taylor, 1956). However, with the marked surge in the development of associative theories of Pavlovian conditioning in the late 1960s and early 1970s (Kamin, 1969; Rescorla & Wagner, 1972) came an attempt

to encompass a greater variety of phenomena, including the UCS preexposure effect, in associative terms.

The most viable associative explanation of the UCS preexposure phenomenon obtained in human eyelid conditioning, rabbit eyelid conditioning, CER, and CTA learning is context blocking. According to this view, prior exposure to the UCS conditions some stimulus aspect of the experimental environment, thereby blocking subsequent conditioning of a nominal CS in that same environment. A question that naturally arises is, what can be gained by the study of the UCS preexposure effect if context blocking forms the basis of this phenomenon? In other words, what implications does a context-blocking explanation have for learning theory in general?

First, a context-blocking analysis should delineate the nature of contextual stimuli that can be associated with UCSs and should determine whether responding to these stimuli can be described by the same laws that govern responding to discrete CSs. Although an influential formal model of Pavlovian conditioning (Rescorla & Wagner, 1972) has emphasized the importance of the conditioning context in an abstract sense, there have been few attempts to identify the physical aspects of the context that do become conditioned. In the absence of such attempts, potentially important differences between conditioning paradigms remain obscure. For example, interest in context blocking has called attention to the importance of handling and injection cues in the analysis of learning established in the CTA paradigm (Braveman, 1978; Rudy, Iwens, & Best, 1977; Willner, 1978). Such stimuli, for example, the handling and injection procedure that occurs between the presentation of the nominal CS and the occurrence of the UCR, have no obvious analogues in the CER or eyelid conditioning paradigms.

A context-blocking view may also provide important information concerning common elements of the UCRs elicited by different UCSs and the role of such responses in the conditioning process. Such a notion was anticipated in Konorski's (1967) discussion of the role of preparatory CRs in the conditioning process. According to Konorski, preparatory or diffuse emotional responses elicited by UCSs must first be conditioned to a nominal CS before a consummatory or specific CR can be established. Information about the similarities among preparatory responses based on different UCSs could be derived from outcomes of experiments on cross-UCS preexposure effects.

Other associative accounts of the UCS preexposure phenomenon, such as controllability of the UCS (Vogel, Note 2), remain to be more fully explored. Research based on the UCS-controllability hypothesis may establish vital links between the learned helplessness phenomenon and the UCS preexposure effect.

Nonassociative accounts of the UCS preexposure phenomenon also warrant thorough investigation because they too have important implications for learning theory. The most viable nonassociative hypothesis of the UCS preexposure effect is that some central adaptation or habituation process occurs during prior exposure to the UCS and reduces the impact of the UCS during subsequent excitatory conditioning. If central habituation proves to form the basis of the UCS preexposure effect in a given paradigm, then one should consider the possibility that a substantial amount of habituation occurs in any Pavlovian conditioning procedure that repeatedly presents the UCS. The opponent-process model (Solomon & Corbit, 1974) makes an analogous prediction, although on the basis of cancellation of a constant *a* process by a gradually increasing *b* process. In any case, the role of *b* processes in habituation of various responses is a worthwhile area for further study.

One widely studied paradigm in Pavlovian conditioning, called *blocking* (Kamin, 1969), refers to the case in which prior conditioning to one stimulus (A) markedly attenuates conditioning to an added stimulus (B) in the compound stimulus AB, as long as the UCS remains unchanged. A common assumption in all explanations of blocking (Mackintosh, 1975; Rescorla & Wagner, 1972) is that to the extent that the added stimulus, B, provides no new information about the

occurrence of the UCS (beyond that provided by Stimulus A), Stimulus B will not condition. However, if it is assumed that habituation occurs to the UCS during Stimulus-A training, then this reduced emotional responsiveness may in part be responsible for the failure to condition Stimulus B, independent of the presence of Stimulus A in either phase. This notion suggests that a group of animals receiving reinforced presentations of Stimulus C followed by reinforced presentations of the compound stimulus AB (C+/AB+) would show blocking relative to a group receiving no treatment and then AB+ training. A C+/AB+ control group is not typically incorporated in studies of the blocking phenomenon.

A habituation account of the UCS preexposure phenomenon also bears on the *unblocking* effect. It is known that unblocking, or conditioning of the added stimulus (B) in the compound AB, occurs following prior conditioning of Stimulus A when the intensity of the UCS used to reinforce the compound is increased (Kamin, 1969) relative to the intensity of the UCS used to condition Stimulus A. Similarly, Dickinson, Hall, and Mackintosh (1976) showed that unblocking occurs when Stimulus A is reinforced with two shocks and the compound stimulus AB is reinforced with only one shock; there is a relative decrease in the intensity of the UCS. The unblocking effect is typically attributed to the surprise value of the novel UCS used to condition Stimulus AB, although the mechanism of surprise remains unstated. In this regard, habituation to the UCS may provide a mechanism of surprise. If one assumes that some habituation of the UCR to the UCS occurs during conditioning of Stimulus A, the use of a novel UCS during conditioning of Stimulus AB, whether the UCS represents a relative increase or a relative decrease in intensity, may act to temporarily dishabituate the UCR to the UCS (Thompson & Spencer, 1966). This treatment would effectively act to restore the former UCR to the UCS, and depending on how much habituation has occurred and the magnitude of the dishabituation effect, would permit some conditioning of Stimulus B.

A dishabituation account of the unblocking effect also revitalizes the explanation of blocking and unblocking that can be derived from the Rescorla–Wagner (1972) model of conditioning. The occurrence of unblocking when Stimulus A is reinforced with two shocks but Stimulus AB is followed by only one shock (Dickinson et al., 1976) is problematical for this model, which assumes that only a manipulation that increases the overall level of associative strength that the UCS can support permits Stimulus B to condition. Reducing the number of shocks during AB training is not expected to produce an increase in overall level of associative strength. However, if unblocking is attributable to dishabituation, then restoration of a previously habituated UCR to the UCS should act to temporarily increase the overall level of associative strength that the UCS can support and permit Stimulus B to condition according to the general framework specified by the Rescorla–Wagner model.

The evidence presented in this review conclusively supports neither a context-blocking nor a habituation account of the UCS preexposure phenomenon; nor is it necessary to assume that a single mechanism of action is responsible for the phenomenon. Associative and nonassociative factors may both play a role in any given situation, and, indeed, the relative importance of the two may vary across conditioning paradigms. At the present time, we suffer from both a lack of understanding of basic processes and the absence of theoretical models that provide rules for combining associative and nonassociative contributions to performance established by Pavlovian conditioning procedures.

Reference Notes

1. Moore, J. W. *Contextual constraints in Pavlovian inhibitory control.* Paper presented at the meeting of the American Psychological Association, New Orleans, August 1974.
2. Vogel, J. R. *Prior exposure to a drug (US) attenuates learned taste aversion.* Paper presented at the meeting of the Psychonomic Society, Boston, November 1974.
3. Goudie, A. J., Thornton, E. W., & Marsh, N. W. A. *Effects of drug experience on subsequent conditioning of drug-induced taste aversions: An evaluation of associative theories.* Unpublished

manuscript, Liverpool University, Liverpool, England, 1977.

4. Zellner, D. A., & Riley, A. L. *Attenuation of food aversions by signalled and unsignalled US pre-exposure.* Paper presented at the meeting of the Eastern Psychological Association, Washington, D.C., March 1978.
5. Gamzu, E. *Pre-exposure to unconditioned stimulus alone may eliminate taste-aversions.* Paper presented at the meeting of the Psychonomic Society, Boston, November 1974.
6. Braveman, N. S. *Conditioned drug states and the treatment pre-exposure effect in taste aversion learning.* Unpublished manuscript, Memorial University, St. John, Canada, 1977.
7. Hinson, R. E., & Siegel, S. *The mechanism of the UCS pre-exposure effect.* Paper presented at the meeting of the Psychonomic Society, Washington, D.C., November 1977.
8. Poulos, C. X., & Cappell, H. *An associative analysis of pretreatment effect in gustatory conditioning by drugs.* Paper presented at the meeting of the American Psychological Association, San Francisco, August 1977.
9. Batson, J. D., & Best P. J. *Blocking of taste aversion learning by environmental cues: The illness preexposure effect is an associative process.* Paper presented at the meeting of the Eastern Psychological Association, Washington, D.C., March 1978.

References

Ader, R. Conditioned adrenocortical steroid elevations in the rat. *Journal of Comparative and Physiological Psychology,* 1977, *90,* 1156–1163.

Amit, Z., & Baum, M. Comment on the increased resistance-to-extinction of an avoidance response. *Psychological Reports,* 1970, *27,* 310.

Baker, A. G. *Rats learn that events, be they stimuli or responses, bear no relation to one another.* Unpublished doctoral dissertation, Dalhousie University, 1974.

Bassett, J. R., Cairncross, K. D., & King, M. G. Parameters of novelty, shock predictability and response contingency in corticosterone release in the rat. *Physiology & Behavior,* 1973, *10,* 901–907.

Berger, B. D., Wise, C. D., & Stein, L. Area postrema damage and bait shyness. *Journal of Comparative and Physiological Psychology,* 1973, *82,* 475–479.

Berman, R. F., & Cannon, D. S. The effect of prior ethanol experience on ethanol-induced saccharin aversion. *Physiology & Behavior,* 1974, *12,* 1041–1044.

Bolles, R. C., & Seelbach, S. E. Punishing and reinforcing effects of noise onset and termination for different responses. *Journal of Comparative and Physiological Psychology,* 1964, *58,* 127–131.

Brady, J. V. Emotion and sensitivity of psychoendocrine systems. In D. C. Glass (Ed.), *Neurophysiology and emotion.* New York: Rockefeller University Press, 1967.

Braveman, N. S. Formation of taste aversions in rats following prior exposure to sickness. *Learning and Motivation,* 1975, *6,* 512–534.

Braveman, N. S. What studies on pre-exposure to pharmacological agents tell us about the nature of the aversion-inducing treatment. In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection.* Waco, Tex.: Baylor University Press, 1977.

Braveman, N. S. The role of handling cues in the treatment preexposure effect in taste aversion learning. *Bulletin of the Psychonomic Society,* 1978, *12,* 74–76.

Brimer, C. J., & Kamin, L. J. Disinhibition, habituation, sensitization, and the conditioned emotional response. *Journal of Comparative and Physiological Psychology,* 1963, *56,* 508–516.

Brookshire, K. H., & Brackbill, R. M. Formation and retention of conditioned taste aversions and UCS habituation. *Bulletin of the Psychonomic Society,* 1976, *7,* 125–128.

Cannon, D. S., Baker, T. B., & Berman, R. F. Taste aversion disruption by drug pretreatment: Dissociative and drug-specific effects. *Pharmacology, Biochemistry and Behavior,* 1977, *6,* 93–100.

Cannon, D. S., Berman, R. F., Baker, T. B., & Atkinson, C. A. Effect of preconditioning unconditioned stimulus experience on learned taste aversions. *Journal of Experimental Psychology: Animal Behavior Processes,* 1975, *1,* 270–284.

Cappell, H., & LeBlanc, A. E. Conditioned aversion by amphetamine: Rates of acquisition and loss of attenuating effects of prior exposure. *Psychopharmacology,* 1975, *43,* 157–162.

Cappell, H., & LeBlanc, A. E. Parametric investigations of the effects of prior exposure to amphetamine and morphine on conditioned gustatory aversion. *Psychopharmacology,* 1977, *51,* 265–271.

Cappell, H., LeBlanc, A. E., & Herling, S. Modification of the punishing effects of psychoactive drugs in rats by previous drug experience. *Journal of Comparative and Physiological Psychology,* 1975, *89,* 347–356.

Deutsch, J. A., & Eisner, A. Ethanol self-administration in the rat induced by forced drinking of ethanol. *Behavioral Biology,* 1977, *20,* 81–90.

DeWied, D. Influence of the anterior pituitary on avoidance learning and escape behavior. *American Journal of Physiology,* 1964, *207,* 255–259.

Dickinson, A., Hall, G., & Mackintosh, N. J. Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes,* 1976, *2,* 313–322.

Domjan, M., & Best, M. R. Paradoxical effects of proximal unconditioned stimulus preexposure: Interference with and conditioning of a taste aversion. *Journal of Experimental Psychology: Animal Behavior Processes,* 1977, *3,* 310–321.

Elkins, R. L. Bait-shyness acquisition and resistance to extinction as functions of US exposure prior to conditioning. *Physiological Psychology,* 1974, *2,* 341–343.

Elsmore, T. F. Saccharine aversion induced by delta-9-tetrahydrocannabinol: Effects of repeated doses

prior to pairing with saccharine. *Proceedings of the 80th Annual Convention of the American Psychological Association*, 1972, *7*, 817–818. (Summary)

Estes, W. K., & Skinner, B. F. Some quantitative properties of anxiety. *Journal of Experimental Psychology*, 1941, *29*, 390–400.

Feldman, J. M. Blocking as a function of added cue intensity. *Animal Learning & Behavior*, 1975, *3*, 98–102.

Gamzu, E. The multifaceted nature of the taste-aversion-inducing agents: Is there a single common factor? In L. M. Barker, M. R. Best, & M. Domjan (Eds.), *Learning mechanisms in food selection*. Waco, Tex.: Baylor University Press, 1977.

Goudie, A. J., & Thornton, E. W. Effects of drug experience on drug induced conditioned taste aversions: Studies with amphetamine and fenfluramine. *Psychopharmacologia*, 1975, *44*, 77–82.

Goudie, A. J., Thornton, E. W., & Wheeler, T. J. Drug pretreatment effects in drug-induced conditioned taste aversions: Effects of drug dose and duration of pretreatment. *Pharmacology, Biochemistry and Behavior*, 1976, *4*, 629–633.

Hall, G., Mackintosh, N. J., Goodall, G., & dal Martello, M. Loss of control by a less valid or by a less salient stimulus compounded with a better predictor of reinforcement. *Learning and Motivation*, 1977, *8*, 145–149.

Hobson, G. N. Effects of UCS adaptation upon conditioning in low and high anxiety men and women. *Journal of Experimental Psychology*, 1968, *76*, 360–363.

Holman, E. W. The effect of drug habituation before and after taste aversion learning in rats. *Animal Learning & Behavior*, 1976, *4*, 329–332.

Kamin, L. J. Apparent adaptation effects in the acquisition of a conditioned emotional response. *Canadian Journal of Psychology*, 1961, *15*, 176–188.

Kamin, L. J. Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior*. New York: Appleton-Century-Crofts, 1969.

Kimble, G. A., & Dufort, R. H. The associative factor in eyelid conditioning. *Journal of Experimental Psychology*, 1956, *52*, 386–391.

Konorski, J. *Integrative activity of the brain*. Chicago: University of Chicago Press, 1967.

Kremer, E. F. Truly random and traditional control procedures in CER conditioning in the rat. *Journal of Comparative and Physiological Psychology*, 1971, *76*, 441–448.

LeBlanc, A. E., & Cappell, H. Attenuation of punishing effects of morphine and amphetamine by chronic prior treatment. *Journal of Comparative and Physiological Psychology*, 1974, *87*, 691–698.

Mackintosh, N. J. Stimulus selection in learning to ignore stimuli that predict no change in reinforcement. In R. A. Hinde & J. S. Hinde (Eds.), *Constraints on learning*. London: Academic Press, 1973.

Mackintosh, N. J. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 1975, *82*, 276–298.

Maier, S. F., & Seligman, M. E. P. Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General*, 1976, *105*, 3–46.

Mikulka, P. J., Leard, B., & Klein, S. B. Illness-alone exposure as a source of interference with acquisition and retention of a taste aversion. *Journal of Experimental Psychology: Animal Behavior Processes*, 1977, *3*, 189–200.

Milulaj, L., & Mitro, A. Endocrine functions during adaptation to stress. *Advances in Experimental Medicine and Biology*, 1972, *33*, 631–638.

Mis, R. W., & Moore, J. W. Effects of preacquisition UCS exposure on classical conditioning of the rabbit's nictitating membrane response. *Learning and Motivation*, 1973, *4*, 108–114.

Munson, P. L. Effects of morphine and related drugs on the corticotrophin (ACTH)-stress reaction. *Progress in Brain Research*, 1973, *39*, 361–372.

Parker, L. F., Failor, A., & Weidman, K. Conditional preferences in the rat with an unnatural need state: Morphine withdrawal. *Journal of Comparative and Physiological Psychology*, 1973, *82*, 294–300.

Pearce, J. M., & Dickinson, A. Pavlovian counterconditioning: Changing the suppressive properties of shock by association with food. *Journal of Experimental Psychology: Animal Behavior Processes*, 1975, *1*, 170–177.

Peck, J. H., & Ader, R. Illness-induced taste aversion under states of deprivation and satiation. *Animal Learning & Behavior*, 1974, *2*, 6–8.

Plotkin, H. C., & Oakley, D. A. Backward conditioning in the rabbit (*Oryctolagus cuniculus*). *Journal of Comparative and Physiological Psychology*, 1975, *88*, 586–590.

Randich, A. *Facilitation and attenuation of the acquisition of a conditioned emotional response by prior exposure to the unconditioned stimulus: An explanation based on the opponent-process theory of acquired motivation*. Unpublished doctoral dissertation, Dalhousie University, 1978.

Rescorla, R. A. Pavlovian conditioning and its proper control procedures. *Psychological Review*, 1967, *74*, 71–80.

Rescorla, R. A. Effect of US habituation following conditioning. *Journal of Comparative and Physiological Psychology*, 1973, *82*, 137–143.

Rescorla, R. A. Effect of inflation of the unconditioned stimulus value following conditioning. *Journal of Comparative and Physiological Psychology*, 1974, *86*, 101–106.

Rescorla, R. A., & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current theory and research*. New York: Appleton-Century-Crofts, 1972.

Revusky, S., Parker, L. A., Coombes, J., & Coombes, S. Rat data which suggest alcoholic beverages

should be swallowed during chemical aversion therapy, not just tasted. *Behavior Research and Therapy,* 1976, *14,* 189–194.

Riley, A. L., Jacobs, W. J., & LoLordo, V. M. Drug exposure and the acquisition and retention of a conditioned taste aversion. *Journal of Comparative and Physiological Psychology,* 1976, *90,* 799–807.

Rudy, J. W., Iwens, J., & Best, P. J. Pairing novel exteroceptive cues and illness reduces illness-induced taste aversion. *Journal of Experimental Psychology: Animal Behavior Processes,* 1977, *3,* 14–25.

Rudy, J. W., Rosenberg, L., & Sandell, J. H. Disruption of a taste familiarity effect by novel exteroceptive stimulation. *Journal of Experimental Psychology: Animal Behavior Processes,* 1977, *3,* 26–36.

Siegel, S., & Domjan, M. Backward conditioning as an inhibitory procedure. *Learning and Motivation,* 1971, *2,* 1–11.

Solomon, R. L. An opponent-process theory of acquired motivation: IV. The affective dynamics of addiction. In M. E. P. Seligman & J. Maser (Eds.), *Laboratory models for psychopathology.* San Francisco: Freeman, 1977.

Solomon, R. L., & Corbit, J. D. An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychological Review,* 1974, *81,* 119–145.

Starr, M. D. *Imprinting in newly-hatched ducklings: Factors influencing the development of separation-induced distress calling.* Unpublished doctoral dissertation, University of Pennsylvania, 1976.

Taylor, J. A. Level of conditioning and intensity of the adaptation stimulus. *Journal of Experimental Psychology,* 1956, *51,* 127–130.

Terry, W. S. Effects of priming unconditioned stimulus representation in short-term memory on Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes,* 1976, *2,* 354–369.

Thompson, R. F., & Spencer, W. A. Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review,* 1966, *73,* 16–43.

Tomie, A. Interference with autoshaping by prior context conditioning. *Journal of Experimental Psychology: Animal Behavior Processes,* 1976, *2,* 323–334.

Wagner, A. R. Priming in STM: An information processing mechanism for self-generated or retrieval-generated depression in performance. In T. J. Tighe & R. N. Leaton (Eds.), *Habituation: Perspectives from child development, animal behavior, and neurophysiology.* Hillsdale, N.J.: Erlbaum, 1976.

Weiss, J. M., McEwen, B. S., Silva, M. T., & Kalkut, M. Pituitary-adrenal alterations and fear responding. *American Journal of Physiology,* 1970, *218,* 864–868.

Whaley, D. L., Scarborough, B. B., & Reichard, S. M. Traumatic shock, x-irradiation and avoidance behavior. *Physiology & Behavior,* 1966, *1,* 93–95.

Willner, J. A. Blocking of a taste aversion by prior pairings of exteroceptive stimuli with illness. *Learning and Motivation,* 1978, *9,* 125–140.

Wise, R. A., Yokel, R. A., & DeWit, H. Both positive reinforcement and conditioned aversion from amphetamine and from apomorphine in rats. *Science,* 1976, *191,* 1273–1274.

Received January 23, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 3, 549–573

Symbolic Interactionist View of Self-Concept: Through the Looking Glass Darkly

J. Sidney Shrauger and Thomas J. Schoeneman
State University of New York at Buffalo

Research on the relationship between self-perceptions and evaluations from other people is reviewed. Studies of naturalistic interactions indicate that people's self-perceptions agree substantially with the way they perceive themselves as being viewed by others. However, there is no consistent agreement between people's self-perceptions and how they are actually viewed by others. There is no clear indication that self-evaluations are influenced by the feedback received from others in naturally occurring situations. When feedback from others is manipulated experimentally, self-perceptions are usually changed. However, methodological limitations such as the questionable external validity and strong demand characteristics of the experimental situations employed make the significance of these findings unclear. The available evidence is examined within a framework that considers the transmission, processing, and evaluation of judgments from others. Other means by which interaction may influence self-perceptions aside from direct evaluative feedback are considered.

O wad some power the giftie gie us
To see oursels as others see us!
Robert Burns, *To a Louse*

Burns's couplet expresses a concern about self-knowledge and its origins that is ancient and contemporary. Recently, a resurgence of interest in the self has flourished in many areas of psychology, especially in psychotherapeutic formulations that view cognitions about oneself as vital mediators in the maintenance and modification of behavior and in social psychological theories involving attribution, cognitive dissonance, and self-awareness. Understanding how attitudes about the self are developed and maintained has thus become increasingly important.

When people are asked how they know that they possess certain characteristics, a typical answer is that they have learned about them from other people. A more formal theoretical statement of this view has been articulated by the influential school of thought known as symbolic interactionism. This theory proffers the idea of a "looking glass self" and asserts that one's self-concept is a reflection of one's perceptions about how one appears to others. This assertion has received widespread professional acceptance and is intoned with catechistic regularity in many leading texts on social behavior (e.g., Raven & Rubin, 1976; D. J. Schneider, 1976; Secord & Backman, 1974).

Social philosophers and psychologists of the late 19th century such as Peirce (1868), James (1890), and Baldwin (1897) were precursors of symbolic interactionism in their emphasis on the self as a product and reflection of social life (Gordon & Gergen, 1968; Ziller, 1973). Cooley (1902), generally credited as the first interactionist, developed the idea of the looking glass self. He posited that the self is inseparable from social life and necessarily involves some reference to others. This process of social reference results in the looking glass self: "A self idea of this sort seems to have three principal elements: the imagination of our appearance to the other person; the imagination of his

---

The authors appreciate the comments of an anonymous reviewer on an earlier draft of this article.

Requests for reprints should be sent to J. Sidney Shrauger, Department of Psychology, State University of New York, 4230 Ridge Lea Road, Buffalo, New York 14226.

judgment of that appearance, and some sort of self-feeling, such as pride or mortification" (Cooley, 1902, p. 152). According to Cooley, from early childhood our concepts of self develop from seeing how others respond to us: "In the presence of one whom we feel to be of importance, there is a tendency to enter into and adopt, by sympathy, his judgment of ourself" (p. 175). Mead (1934), the major theorist of symbolic interactionism, amplified and expanded the view of the self as a product of social interaction: "The individual experiences himself as such, not directly, but only indirectly, from the particular standpoints of other individuals of the same social group, or from the generalized standpoint of the social group as a whole to which he belongs" (p. 138). Essential to the genesis of the self is the development of the ability to take the role of the other and particularly to perceive the attitude of the other toward the perceiver. Mead's looking glass self is reflective not only of significant others, as Cooley suggested, but of a generalized other, that is, one's whole sociocultural environment. More recently, Kinch (1963) has summarized and systematized symbolic interactionist self theory by noting that it basically involves an interrelation of four components: our self-concept, our perception of others' attitudes and responses to us, the actual attitudes and responses of others to us, and our behavior.

In recent years, self theories have been proposed that do not insist on the primacy of social others as sources of information about the self. Bem (1967, 1972) has asserted that self-perception is a special case of person perception:

> Self-descriptive attitude statements can be based on the individual's observations of his own overt behavior and the external stimulus conditions under which it occurs. . . . As such, his statements are functionally similar to those that any outside observer could make about him. (1967, pp. 185–186)

Jones and Nisbett (1971) have qualified Bem's analysis somewhat by proposing that "actors tend to attribute the causes of their behavior to stimuli inherent in the situation, while observers tend to attribute behavior to stable dispositions of the actor" (p. 93). Duval and Wicklund's (1972) objective self-awareness theory also emphasizes the potential of the individual for active self-appraisal. Objective self-awareness is a state of consciousness in which attention is focused inward on the self, making the individual an object to his or her consciousness. The assumption that self-awareness is dependent on the imagination of another's views is minimized. Although these self-perception theories have stimulated considerable research, the initial justification for each view was mainly on theoretical rather than empirical grounds. Thus, some attention is given to the relevance of the data presented here to self-perception theories, although the main objective is to evaluate the evidence relevant to the looking glass self.

Information concerning the looking glass self derives from several lines of inquiry, not all of them explicitly related to this theory. Even work that has been done within the framework of symbolic interactionism suffers from a severe case of "ahistoricity," so that there is little sense of cumulative development of information. This article attempts to examine thoroughly the studies done under the auspices of symbolic interactionism. An exhaustive review of relevant studies outside of this framework cannot be claimed, however, since these come from many divergent bodies of literature.

The research presented is divided into two sections. First, studies are reported that examine feedback given in uncontrolled, naturally occurring interactions. Next, investigations of the effects of controlled feedback in structured situations are considered, with attention given to work in which feedback is purportedly based either on objective information or on more subjective judgments. Some restrictions on the types of research reviewed here should be noted. The main dependent variable examined is expressed self-perceptions. Studies exploring the impact of self-relevant feedback on other aspects of behavior are typically not covered, since it is debatable whether such changes are necessarily mediated by changes in self-perceptions. Also, although it may be argued that studies of attitude change on any topic involve some implied reappraisal of self-evalu-

Table 1
Investigations of the Relationship Between Subjects' Self-Descriptions and Subjects' Perceptions of How Others Describe Them

Study	Analysis	Subjects	Assessment dimension	Significant others	Congruence between self and perceived others' evaluation?[a]
Davidson & Lang (1960)	Correlation	203 fourth–sixth graders	Self-concept	Teacher	Yes (r = .82)
Fey (1955)	Correlation	58 3rd year medical students	Self-acceptance	Peers	Yes (r = .71)
Goodman (1973)	Correlation	185 fourth–sixth graders	Self-concept	Peers	Yes (r = .37)
Goslin (1962)	Comparison	187 seventh and eighth graders	Personality traits	Peers	Accepted subjects: yes; rejected subjects: no
Jourard & Remy (1955)	Correlation	99 undergraduates	Perception of self and body	Mothers, fathers	Yes ($.56 \leq r \leq .77$)
Kemper (1966)	Count of significant correlations	256 business men (M age = 40 years)	Self-description	Wife, boss, colleague, father	Yes (with wife > with boss > with colleague > with father: all rs $\leq .22$)
Miyamoto & Dornbusch (1956)	Count of supporting results	195 undergraduates	Personality traits	Peers, generalized other	Yes
Orpen & Bush (1974)	Correlation	14 males, 17 years old	Responsibility, sociability	Peers	Yes ($.49 \leq r \leq .80$)
Quarantelli & Cooper (1966)	Nonstatistical comparison	1,012 1st to 2nd year dental students	Self-rating on dental student – dentist continuum	Peers, instructors, wives, generalized others	Yes (high self-raters > low in mean perceived response of others)
Reeder, Donohue, & Biblarz (1960)	Nonstatistical comparison	54 enlisted military personnel	Leadership, work ability	Peers	Yes (high self-raters > medium > low in estimated group rating)
Swanson (1969)	Comparison of correlations	11 emotionally disturbed, 35 learning disabled, 35 normal children, 6–12 years old	Self-worth, self-acceptance	Parents	Emotionally disturbed: yes; learning disabled and normals: no
Teichman (1972)	Correlation	50 delinquent, 25 nondelinquent boys	Self-concept	Parents	Nondelinquents: yes ($.78 \leq r \leq .85$); delinquents: no
Walhood & Klopfer (1971)	Correlation	13 graduate students	Love, dominance	Peers	Yes ($.60 \leq r \leq .93$)

[a] For all rs, $p < .05$.

ations, the focus here is limited to changes in attitudes about the self, since there is evidence that reactions to feedback about the self differ from those about other attitudes (e.g., Eagly, 1967). A final restriction involves the area of self-presentation. Expressing one's self-perceptions in any public fashion inevitably has some potential instrumental value, and numerous investigations have focused on the functional impact of such self-statements. These studies, however, address issues that are not central to our discussion. The focus of this article is on investigations in which self-statements are perceived as fairly accurate estimates of the individual's actual attitudes and external incentives to a particular type of self-presentation are minimized.

Naturalistic Studies

Many investigations have sought support for the idea of the looking glass self in naturally occurring interactions. One group of studies has focused on the proposition that individuals' self-perceptions should be highly congruent with the way they see themselves as being perceived by others. Table 1 shows that these studies vary widely along a number of different dimensions. Most analyses were correlational, some involved statistical comparisons, and some of the earlier studies relied on nonstatistical "eyeballing" of the data (e.g., Miyamoto & Dornbusch, 1956; Quarantelli & Cooper, 1966; Reeder, Donohue, & Biblarz, 1960). Samples have been drawn from all levels of the educational system and from a variety of other populations. Evaluations by self and others have most often centered on global measures of self-concept, although some investigations have examined more specific aspects of personality and behavior. Overall, these studies show modest to strong correlations between individuals' perceptions of themselves and the way they assume others perceive them. Nonsignificant relationships have occurred in situations in which deviant groups, such as delinquents (Teichman, 1972), learning disabled students (Swanson, 1969) and sociometrically rejected students (Goslin, 1962), have been studied. The only exception to this pattern is Swanson's finding that for 11 emotionally disturbed children there was congruence between self-acceptance and perceived parental acceptance and that for 35 normal children this congruence was absent.

In addition to postulating concordance between self-evaluation and the perceived evaluations of *significant* others, Mead (1934) contended that self-concept is reflective of the perceived evaluation of a *generalized* other. Relatively few studies have examined this facet of symbolic interactionism. There is some evidence that individuals' self-perceptions are similar to their perceptions of how they are viewed by others in general (Miyamoto & Dornbusch, 1956; Quarantelli & Cooper, 1966; Reeder et al., 1960). The evidence on whether self-perceptions are more strongly related to the perceived impressions of specific others or to the perceived impressions of the generalized other, however, is contradictory (Miyamoto & Dornbusch, 1956; Quarantelli & Cooper, 1966).

The demonstration of a relationship between people's self-perceptions and how they feel others see them is not sufficient in validating the symbolic interactionist position. It is necessary, in addition, to demonstrate congruence between (a) self-perceptions and others' actual perceptions of the person and (b) perceived other-evaluations and actual other-evaluations. A large number of studies have examined the former relationship; they are summarized in Table 2. Although many of these studies are of questionable statistical and conceptual significance (Wylie, 1974), the overall pattern of the conclusions drawn by these investigations suggests much less agreement between self-judgments and actual judgments by others than between self-judgments and perceived judgments. Approximately half the studies reviewed show no significant correlations between self-perceptions and others' actual evaluations. The majority of the remaining investigations have reported either significant but low correlations or ambiguous results. There are no easily distinguishable factors that account for the presence or absence of positive associations. A wide range of subjects and evalu-

(*text continued on page 558*)

Table 2
Investigations of the Relationship Between Subjects' Self-Descriptions and Actual Descriptions of Them by Significant Others

Study	Analysis	Subjects	Assessment dimension	Significant others	Congruence between self and others' evaluations?[a]
Alberti (1971)	Correlation	656 first–third graders	Performance in role of student	Peers, teachers	With peers: no; with teachers: yes
Amatora (1956)	Correlation	400 fourth–eighth graders	22 personality traits	Peers	Yes: for boys, $.15 \leq r \leq .67$ for 19 out of 22 traits; for girls, $.15 \leq r \leq .62$ for 20 out of 22 traits
Bishop (1971)	Correlation	25 graduate students	Counseling effectiveness	Clients, supervisors	With clients: no; with supervisors: yes ($r = .41$)
Bledsoe & Wiggins (1973)	Comparison	100 ninth graders, 200 parents	Self-image	Parents	No (parents' evaluations > adolescents' self-evaluations)
Brams (1961)	Correlation	27 graduate students	Effective communication in counseling	Peers, supervisors	No
Breslin (1968)	Correlation	28 handicapped, 10–16 years old	Self-concept	Peers	No
Buckley (1970)	Comparison	22 student teachers	Self-attitude	Students	No (high self-raters equaled low in ratings by students)
Burke (1969)	% agreement	113 undergraduates	Final grade	Peers, teachers	With peers: 70%; with teachers: 60%
Carroll (1952)	Correlation	125 army enlisted men	5 personality traits	Peers	Yes ($.29 \leq r \leq .56$)
Cogan, Conklin, & Hollingworth (1915)	Average deviation of self-rankings from median ranking of subject by others	25 junior, 25 senior coeds	9 personality traits	Peers	No: mean of average deviations on 9 traits was 6.1 places
Douce (1970)	Correlation	60 female high school students	Self-esteem	Peers	No
Eisenmann & Robinson (1968)	Correlation	17 institutionalized physically disabled, 30–60 years old	Creativity	Peers	Yes ($r = .74$)
Fey (1955)	Correlation	58 3rd year medical students	Self-acceptance	Peers	No

(*table continued*)

Table 2 (*continued*)

Study	Analysis	Subjects	Assessment dimension	Significant others	Congruence between self and others' evaluations?[a]
Friedsam & Martin (1963)	Correlation, chi-square	87 medical outpatients over 50 years old	Rating of health	Physicians	Yes: corrected coefficient of contingency = .33; $\chi^2 = 5.00$, $p < .05$, for table of favorable–unfavorable self-ratings × favorable–unfavorable physician ratings of health
Goldings (1954)	Rank order correlation	20 male undergraduates	Happiness	5 "experienced judges" (experimenters who had tested subjects)	Ambiguous: for 2 judges, $r = .45$ and $.64$ ($p < .05$); for other 3 judges, p not significant (for all 5 judges, average $r = .38$)
Goodman (1973)	Correlation	185 fourth–sixth graders	Self-concept	Peers	Yes ($r = .20$)
Goslin (1962)	Comparison	187 seventh and eighth graders	Personality traits	Peers	Accepted subjects: yes; rejected subjects: no
Gray & Gaier (1974)	Correlation	7 12th-grade females, 14 best friends, 14 parents	Positive and negative traits	Peers, Parents	Yes (with parents, $r = .74$; with peers, $r = .76$)
Green (1948)	Comparison	23 Egyptian male graduate students and 23 Egyptian, British, and Greek female undergraduates	Leadership ability	Peers	No: 20 out of 23 males, 16 out of 23 females overestimated own rank as compared with group's ranking of him or her
Hamilton (1969)	Correlation	70 fraternity members	Self-esteem dominance, dogmatism	Peers	No
Hase & Goldberg (1967)	Correlation	201 1st year coed-dormitory residents	5 personality traits	Peers	Yes ($.23 \leq r \leq .56$)
Helper (1958)	Correlation	53 eighth and ninth graders	Favorableness, acceptance	Parents	With fathers: yes ($.26 \leq r \leq .44$); with mothers: no
Horowitz (1962)	Correlation	111 fourth–sixth graders	Self-concept	Peers (same sex)	Fourth-grade girls: yes ($r = .59$); all others: no

Table 2 (*continued*)

Study	Analysis	Subjects	Assessment dimension	Significant others	Congruence between self and others' evaluations?[a]
Israel (1958)	Comparison of level of self-evaluation of subjects evaluated by group as high, middle, low No. of subjects whose comparison of self versus others agreed with the others' comparisons	29 student nurses	Intelligence, leadership, orderliness, appearance	Peers	Yes: for intelligence and leadership only, subjects highly evaluated by group had higher self-evaluations No: few subjects exceeded chance accuracy in their comparisons of self versus others versus the same comparisons made by the others
Jansen, Robb, & Bonk (1973)	Factor analysis	173 graduate students	Counseling competence	Peers	No (self-ratings > peer ratings)
Jorgenson (1967)	Correlation	400 third graders	Personality traits	Peers, teachers	No
Kelman & Parloff (1957)	Rank order correlations	7 male, 8 female neurotic group-therapy patients	Behaviors in group therapy	3 observers (psychologists and social worker)	No (both before and after therapy)
Klimoski & London (1974)	Correlation, factor analysis	133 nurses	Job performance	Peers, supervisors	No
Lomont (1966)	Comparison	64 sorority, 72 fraternity members	Dominance–submission, love–hate	Peers	Yes: no significant differences between means of subjects' self-ratings and means of peer ratings
Mayo & Manning (1961)	Correlation	196 naval recruits	Effort in aviation course	Peers	Yes (r = .28)
McConnell (1959)	Comparison	137 third and fourth graders	Social acceptance	Peers	No: less than 50% females, about 33% males judged selves accurately (compared with peer ratings)
McIntyre (1952)	Comparison	224 second semester dormitory freshmen	Self-acceptance	Peers	No (high-accepted subjects equaled low in self-acceptance)

(*table continued*)

Table 2 *(continued)*

Study	Analysis	Subjects	Assessment dimension	Significant others	Congruence between self and others' evaluations?[a]
Miyamoto & Dornbusch (1956)	Count of significant results	195 undergraduates	Personality traits	Peers	Yes
Mote (1967)	Correlation	157 fifth and sixth graders, mothers	Self-concept	Mothers	Mother's satisfaction with child learning: yes; with child behavior: no
Orpen & Bush (1974)	Correlation	14 males, 17 years old	Responsibility, sociability	Peers	No
Perkins (1958)	Correlation	48 fourth–sixth graders	Self-evaluation	Teachers	Yes (r = .41)
Phillips (1963)	Correlation	96 third and 96 sixth graders	Social characteristics, school achievement inventory	Peers, teachers	Third graders: no; sixth graders: yes (with peers, r = .40; with teachers, r = .57)
Powell (1948)	Rank order correlations	140 coed-dormitory residents	Adjustment	Dorm advisors, peers	No: *Mdn ps* (from 8 dormitory corridors) $\leq$ .24
Reeder, Donohue, & Biblarz (1960)	Nonstatistical comparison	54 enlisted military personnel	Leadership, work ability	Peers	Low self-raters: yes; high and medium self-raters: no
Reese (1961)	Comparison	408 fourth, sixth, eighth graders	Self-evaluation	Peers	Curvilinear relationship: medium self-concept > high > low in amount of acceptance by others
Rokeach (1945)	No. overestimates versus underestimates (self versus average other ratings)	134 female undergraduates	Physical beauty	Peers	No: 72% of subjects overestimated own beauty
B. Schneider (1970)	Correlation, factor analysis	240 male undergraduates	Group leadership behaviors	Peers	Yes ($.25 \leq r \leq .51$); factor structure different for self-ratings and peer ratings

Table 2 (*continued*)

Study	Analysis	Subjects	Assessment dimension	Significant others	Congruence between self and others' evaluations?[a]
Scott & Johnson (1972)	Correlation	234 undergraduates	Attitudes toward 14 persons, issues, and so on	Friends	Yes ($.14 \leq r \leq .61$)
	Correlation	50 undergraduates	9 motives (needs)	Friends	Yes, for 7 of 9 motives ($.25 \leq r \leq .55$)
	Correlation	92 institutionalized youthful offenders	Attitudes toward 16 persons, issues, and so on	Work supervisors, counselors	Yes, for 8 of 16 attitudes ($.17 \leq r \leq .59$)
Todorosky (1972)	Correlation	177 sorority members	Self-acceptance	Peers	No
Tschechtelin (1945)	Comparison	1,542 fourth–eighth graders	22 personality traits	Teachers, peers	No: boys tended to underestimate self versus peers and teachers; girls tended to overestimate
Walhood & Klopfer (1971)	Correlation	13 graduate students	Love, dominance	Peers	No
Webb (1955)	Correlation	95 naval aviation cadets	Intelligence	Peers	Yes ($r = .43$)
Werdelin (1969)	Correlation, factor analysis	416 high school students	Classroom behaviors	Peers, teachers	No
Wetzel (cited in Peterson, 1965)	Correlation	72 undergraduates	Adjustment, introversion–extraversion	Peers, parents	Yes ($.24 \leq r \leq .41$)
Winthrop (1959)	Correlation	60 female undergraduates	Adjustment	Closest friend	Yes: for overall adjustment and 4 subscales, $.41 \leq r < .66$

[a] For all *r*s, $p < .05$.

(*text continued from page 552*)

ators were used, and comparisons were made on many attributes, most frequently self-esteem or task competence. Also, a number of studies have shown that perceived reactions of others are closer to self-concept than are actual reactions (Miyamoto & Dornbusch, 1956; Orpen & Bush, 1974; Quarantelli & Cooper, 1966; Sherwood, 1965; Walhood & Klopfer, 1971). The minimal associations between self-perceptions and others' actual evaluations suggest that people do not accurately perceive others' opinions of them, that these opinions minimally influence self-judgments, or, as indicated by a study by Reese (1961), that these two variables may be curvilinearly related, thus explaining why significant linear correlations do not often emerge (Hartup, 1970). Studies assessing degree of influence are infrequent and are discussed below.

The issue of accuracy in perceiving others' opinions has also been examined by the consideration of the relationship between individuals' perceptions of others' views of them and others' actual views. Of the studies assessing this relationship, some show congruence (Ausubel & Schiff, 1955; Ausubel, Schiff, & Gasser, 1952; De Jung & Gardner, 1962), some indicate partial or ambiguous relationships (Goslin, 1962; Israel, 1958; Reeder et al., 1960; Tagiuri, Blake, & Bruner, 1953; Walhood & Klopfer, 1971), and others demonstrate no association (Ausubel, 1955; Fey, 1955; Kelman & Parloff, 1957; Orpen & Bush, 1974). Most of the studies showing congruence have involved judgments of highly evaluative characteristics such as liking by the other person, whereas those showing minimal associations have typically involved more content-specific judgments. Ability to predict peers' liking increases with age, at least from the lower grades through high school (Ausubel & Schiff, 1955; Ausubel et al., 1952; De Jung & Gardner, 1962), reflecting perhaps a more extentive interaction with those judged, more frequent expression of interpersonal preferences, or greater sensitivity to interpersonal cues. Also, whether one is predicting positive or negative feelings may be important; people seem to be better able to predict who likes them best as opposed to who likes them least (Tagiuri et al., 1953). That self-perceptions are consistently more strongly correlated with people's perceptions of how they think others view them than with how others actually view them suggests that the tendency to assume greater similarity between one's own and others' attitudes than actually exists (e.g., Newcomb, 1961) extends into the area of attitudes toward oneself. Thus, subjects' self-evaluations may be weakly related to others' opinions of them because they frequently do not know what others' opinions are.

Since the studies reported thus far show no direction of causality or change over time, it is impossible to decide whether the actual or perceived evaluations of people by others are a cause or effect of how they perceive themselves. If one is to infer that others' judgments influence self-perception, assessments must be made at different times to see if self-perceptions change in the direction of others' earlier evaluations. Almost all of the relevant investigations have examined short-term changes in self-evaluation in relation to actual or perceived evaluations by others. Sherwood (1965) had sensitivity training participants rate themselves on a set of bipolar trait scales during the second day of a 2-week program. At the end of the program they rated themselves again, rated how they felt other group members would rate them, and rated other group members on the same set of dimensions. The ratings of a person by others were more similar to his or her self-ratings at the end of the program than to initial self-ratings. Since others' ratings were not obtained at the outset, one cannot infer that they actually influenced self-ratings. Instead, both the subject and other group members may have observed and responded to changes in subjects' presentation of themselves as the sessions continued.

Rosengren (1961), in a study of 10 institutionalized preadolescent boys with emotional disturbances, obtained self-ratings and ratings by peers over a 1-year interval. He found that for the post- as compared with the preratings, self-ratings were more similar

to both subjects' perceptions of others' ratings of them and others' actual ratings of them. Although these subjects did see themselves more similarly to the way they were seen by others, the critical comparison showing that self-ratings in the second evaluation became more similar to others' initial evaluations of them was not made.

The most sophisticated naturalistic investigation to date remains an early study by Manis (1955). Male undergraduates assigned as dormitory roommates rated themselves, their ideal selves, and their roommates at the beginning of a semester and after 6 weeks. Based on sociometric choices at the beginning of the first sessions, a friend and a nonfriend were designated for each subject. Subjects' self-perceptions and their friends' perceptions of them were more similar after their final rating than after their first. The most important finding was that subjects' final self-ratings were more similar to others' initial judgments of them than were their initial self-ratings. Furthermore, others' second ratings of a subject were no more similar to the subject's initial self-perceptions than were their first ratings, suggesting that others' impressions were not substantially influenced by the subject's initial self-evaluation.

Although these data suggest that individuals do change their self-perceptions in the direction of others' opinions about them, methodological limitations make this conclusion equivocal. Most significantly, subjects' self-perceptions changed in the direction of friends' initial judgments of them only when the designated friend had initially described them more favorably than subjects had described themselves. When their designated friend described them less positively than their own self-perceptions, there were no increases in the similarity of their self-descriptions. A friend who views subjects more positively than the subjects view themselves would be likely to reciprocate the subjects' friendship more than someone who views them less favorably than they view themselves. Learning that they have chosen as a friend someone who also likes them may enhance people's feelings of interpersonal perceptiveness and social competence and cause them to raise their self-evaluation.

Even if Manis's results indicate that a friend who describes a peer positively influences the peer's self-perceptions, the nature of the changes generated remains unclear. Subjects may either change the overall favorableness of their self-ratings to more closely match that of their evaluators, or they may change their assessments on specific dimensions so that the pattern of their self-descriptions across dimensions becomes more similar to that of their evaluators. This distinction raises the issue of whether the influence of others' assessments extends beyond the general evaluative level to more specific elements of the dimension being assessed. Perhaps when people are reacting to others' evaluations of them, the principal or even exclusive information that they process is whether they are being perceived in some globally positive, negative, or neutral way.

The only long-term longitudinal study that has been reported involved self-ratings and ratings by peers and teachers of children in the first and second grades who were later reassessed in the fifth and sixth grades (Trickett, 1969). Neither peer nor teacher ratings from the initial assessment were significantly correlated with self-ratings in the second assessment. Children's perceptions of how peers saw them in the initial assessment were uncorrelated with their self-perceptions in the second measurement. Although the author implied some causal influence of others' ratings (particularly those of teachers) on later self-perception, this is difficult to detect in the data. The absence of such an effect is not surprising in light of the fact that by the time subjects were in the later grades they had been exposed to a number of different peers and teachers, whose influence was impossible to gauge.

The numerous naturalistic studies that have been undertaken have not, by and large, contributed substantially to an understanding of the extent to which others' perceptions influence self-judgments. Currently, there is little evidence that in their ongoing social interactions people's views of themselves are shaped by the opinions of others. This is due

primarily to the lack of repeated assessments of self-perceptions and others' perceptions whereby movements of one toward the position of the other could be determined.

Other issues are also important in evaluating the naturalistic data. Many investigations may not have examined situations in which the input of other people was maximal. For instance, most studies have used late adolescents and adults as subjects. If these individuals are in stable life situations, they may be more likely to maintain relatively solidified self-images. The impact of others' opinions could possibly be enhanced and more pronounced if adults were studied in unfamiliar situations in which their norms for self-evaluation and the behavior patterns that they displayed were in a state of flux, as in Manis's (1955) study of incoming college freshmen in dormitories. It also seems likely that younger people are more susceptible to external influence in developing their self-concept than are older individuals.

A final consideration in assessing the work reviewed above concerns the individuals who are sources of feedback and their relationship to the subjects studied. Although peers are the most commonly used and are, in many cases, perhaps the most appropriate sources of evaluations, more attention should be given to the actual degree of interaction between them and the people whose self-perceptions are being assessed. Membership as a peer in a group of students or workers does not necessarily demand that colleagues offer appraisals to one another. For both children and adults, a relatively small number of people may serve as significant sources of evaluative feedback. In most studies it is the researcher who decides who the subjects' significant others are, and in many cases this designation may be off the mark. Investigations that attempt to identify the significant others of a given population (e.g., Denzin, 1966) would be useful preliminary steps in future naturalistic investigations.

Studies of Controlled Feedback From Others

Although researchers have employed a wide range of specific procedures for assessing the role of controlled feedback on judgments of others, most studies have followed one of two paradigms, which differ mainly in the extent to which the evaluator's judgments are based on objective data. In the first type of study the feedback received is purportedly based on tests of personality or competence. Typically, subjects describe themselves on the attributes assessed by the tests, then take the tests, receive feedback about their performance either immediately or within a week or two, and finally reappraise themselves. This procedure has been employed not only in specific efforts to assess the symbolic interactionist position but also in studies examining the effects of change in self-evaluation on other aspects of behavior, with change in self-evaluation often examined principally as a manipulation check. In the second type of study, feedback is based on the subjective impressions of other individuals who have no specific knowledge of objective assessment results. These studies have varied in the extent to which the other person is presented as having expertise in the topics considered.

The most elementary question typically asked in this research is, Will individuals modify their self-descriptions in the direction of the feedback they receive? The most elementary answer is *usually*. Such changes have been shown for numerous populations and for many different attributes, from competence in public speaking (Videbeck, 1960) and physical skills (Haas & Maehr, 1965) to a variety of personality traits (e.g., Backman, Secord, & Pierce, 1963; Binderman, Fretz, Scott, & Abrams, 1972; Cooper & Duncan, 1971; Eagly, 1967; Evans, 1962; Harvey & Clapp, 1965; Harvey, Kelley, & Shapiro, 1957; Regan, Gosselink, Hubsch, & Ulsh, 1975; Shrauger & Lund, 1975; Snyder & Shenkel, 1976; Steiner, 1968). In almost all cases changes in self-perception have been judged by modifications in verbal self-descriptions made immediately following others' evaluations and in the presence of the evaluator.

Although controlled feedback from others typically produces some changes in people's self-descriptions, several factors influence the extent of such changes. These include the

discrepancy of feedback from subjects' self-perceptions, favorableness of feedback, characteristics of the evaluator, consensual validation of the judgments given, and attributes of those evaluated. After these factors have been examined, some general observations on the significance and limitations of studies employing manipulated feedback are considered.

Discrepancy of Feedback From Self-Perceptions

The amount of discrepancy between others' evaluations and one's own self-perceptions has been examined in several studies. Bergin (1962) found that the credibility of feedback influenced the relationship between discrepancy and self-perception changes. With a high-credibility source, increases in discrepancy resulted in greater changes in self-relevant attitudes, whereas for a low-credibility source the tendency was for greater credibility to produce less change. Although not wholly consistent, other results have suggested that when others' evaluations are purportedly based on objective test data, self-perceptions change more as the discrepancy from initial perceptions increases (Binderman et al., 1972; Eagly, 1967; Gerard, 1961; Johnson, 1966). However, Gerard found that this occurred only when subjects felt that the feedback they had received would be made public, and Eagly found that changes increased from low to moderate but not from moderate to high levels of discrepancy. Johnson found a curvilinear trend, with attitude change first increasing with increased discrepancy and then decreasing. In contrast with the findings based on objective test data, when feedback was based on subjective ratings of a personality dimension made by subjects' classmates, changes in self-evaluations were not enhanced by increased discrepancy between their judgments and subjects' initial self-perceptions (Harvey & Clapp, 1965; Harvey et al., 1957). Although many factors may differentiate these studies from one another, they are generally consistent with Bergin's argument about the role of credibility and suggest that for feedback that diverges substantially from one's views to have a strong effect on self-evaluations, it must be perceived as being based on clear objective information.

Favorableness

Several studies have examined amount of change in self-perceptions as a function of feedback favorableness. Some of these studies involved the "Barnum effect," that is, the acceptance of bogus personality feedback (Meehl, 1956). Most such investigations indicate that favorable information is more readily accepted than unfavorable information (Sundberg, 1955; Halperin, Snyder, Shenkel, & Houston, in press; Mosher, 1965; Weisberg, 1970), with a few showing no differential acceptance (Dmitruk, Collins, & Clinger, 1973; Evans, 1962). These studies' significance is questionable, however, since they involved no preassessments of subjects' self-evaluations and may have reflected the greater comparability between positive information and initial self-perception than between negative information and initial self-perceptions.

A few studies have attempted to control for the discrepancy between feedback and initial impressions. Steiner (1968) examined changes in self-ratings on bipolar traits and found that positive feedback produced greater changes than negative feedback, when feedback was based on upper level undergraduates' interpretations of self-report tests. Another study (Snyder & Shenkel, 1976) attempted to control for the "initial truthfulness" of the information evaluated and found no differences in the acceptance of positive or negative feedback given by the graduate student and based on projective tests.

Turning to studies in which feedback was not based on personality test results, we find that most were not designed to assess differences in reactions to equally discrepant positive and negative evaluations (Haas & Maehr, 1965; Jones, Gergen, & Davis, 1962; Maehr, Mensing, & Nafzger, 1962; Papageorgis & McCann, 1965; Videbeck, 1960). Two careful investigations that did examine initial self-perceptions produced inconsistent findings similar to those of Steiner and Sny-

der and Shenkel just discussed. Eagly (1967) found no differential acceptance of feedback from a trained rater with regard to subjects' assertiveness or submissiveness. Harvey and Clapp (1965), however, found that students changed their self-ratings on a set of bipolar adjectives more when they had received positive feedback than when they had received negative feedback from classmates. The evaluators in Eagly's study may have had more legitimized expertise than those in Harvey and Clapp's study, and the same may have been true in the Barnum effect study of Snyder and Shenkel (1976) versus that of Steiner (1968). These results may suggest that subjects are reluctant to accept unflattering information about themselves unless they feel that the source of that information has a particularly strong basis for judgment. The inconsistency of these findings, however, suggests that the differential acceptance of positive versus negative information may depend on a variety of parameters. Eagly and her colleagues have shown, for example, that positive information is readily accepted if the recipients of the information do not expect to be evaluated again (Eagly & Acksen, 1971) and if they have no choice over the information they have received (Eagly & Whitehead, 1972). Other factors such as the strength of the subject's initial self-perceptions and the attributes on which feedback was given may also be relevant here.

Evaluator Characteristics

The most systematic investigation of factors affecting the influence of an information source involved several studies by Webster and Sobieszek (1974), who examined subjects' responses to evaluations of their ability on a perceptual task. Each subject worked with a partner, and both subjects' initial performance was judged by an evaluator whose apparent competence on the task was varied. The impact of the evaluator's assessment on subjects' self-perceptions was not measured directly, but was inferred from the extent to which subjects acquiesced to their partners' judgments on a subsequent set of items. The evaluator's judgments had more effect on rate of acquiescence when the evaluator was presented as very competent as opposed to moderately competent and had no effect when he was presented as incompetent. Manipulation of more general aspects of the evaluator's competence by presenting him to high school subjects as either a college junior or an eighth grader produced no differential changes in acquiescence level.

Other investigators have also examined the effects of manipulating general competence. Whether a test evaluator was a PhD or a counseling practicum student influenced the degree of acceptance of bogus personality feedback if that information was highly discrepant from the subject's initial self-perception but not if it was less discrepant (Binderman et al., 1972). Whether a person received ratings on adjective dimensions from an acquaintance or from a stranger in his or her class had no effect on the degree of change in subsequent self-ratings (Harvey et al., 1957). Although it is difficult to develop generalizations from such scattered findings, these data suggest that others' expertise or competence has an impact on the acceptance of their evaluations only when that competence is specifically relevant to the judgment being made.

Consensual Validation

Another aspect of the credibility of information received involves the extent to which it is validated by others. Presumably, as a larger number of individuals reflect a particular perception to the subject, the likelihood that the subject will incorporate that perception is increased. Following this assumption, Backman et al. (1963) found that bogus personality feedback had less effect on college students' self-ratings as a greater number of significant others were viewed as agreeing with the subject's initial self-perception. The specific relevance of the number of others who hold an opinion is unclear, however, since neither the salience of the dimensions to the subjects themselves nor the strength of their own self-perceptions was assessed. In another study, junior high school boys were given feedback about their physical skills by either one or two experts (Haas &

Maehr, 1965). Initial postfeedback ratings did not differ as a function of the number of raters, but self-ratings made 6 weeks after the experts' judgments showed greater changes on the attributes evaluated for the group judged by two experts. Since there was no condition in which consistent feedback was repeated by a single judge, it is not certain that it was a second person as opposed to a repetition of the communication that was the critical factor in enhancing feedback. This is an important issue, since there is some evidence that the repetition of an evaluation by the same evaluators enhances changes in self-evaluations (Kinch, 1968). Thus, there is no clear evidence that increasing the number of people who make an evaluation enhances the likelihood that it will be accepted.

The consistency of feedback across different evaluators has also been examined. Although there has been some suggestion that people respond more strongly to feedback that is consistent than to feedback that varies from evaluator to evaluator (Sherwood, 1967), other findings offer little support for this view (Kinch, 1968; Sobieszek & Webster, 1973). Because of the wide variation in the methodology of these studies, it is impossible to determine which differences among them account for the inconsistency in findings. However, given the ambiguous nature of these results, plus the fact that multiple and inconsistent evaluations may be frequent in real-life interactions, more careful examination of how evaluative information is combined and integrated seems warranted.

Self-Evaluator Characteristics

There is some evidence that individuals differ in their receptivity to information about themselves. The main characteristic that has been examined in this regard is level of self-esteem, perhaps because most of the work on response to others' feedback has focused on highly evaluative information. There is some consistency in the finding that individuals who have generally low self-esteem are more influenced by negative feedback from others and less by positive feedback than are individuals with high self-esteem. This has been shown even when subjects' initial self-perceptions about the specific attributes evaluated were comparable, and it has been demonstrated for judgments of assertiveness–submissiveness (Eagly, 1967), social sensitivity (Shrauger & Rosenberg, 1970), and several other personality traits (Harvey & Clapp, 1965). The only instance in which such a differential acceptance was not demonstrated was for self-awareness (Shrauger & Lund, 1975).

Studies of other individual differences in recipients have been more episodic. Gerard (1961) found that a self-report measure of susceptibility to social influence predicted degree of change in self-perception, but only when the evaluation from others was supposed to be made public. People who had a less well-developed sense of self or a lower level of ego identity (Erikson, 1956) changed their self-evaluations more following success or failure feedback on an intellectual task than did those at higher levels of ego identity (Marcia, 1967). Harvey, Hunt, and Schroder (1961) reported that levels of concreteness–abstractness in cognitive processes predicted the extent of changes in self-descriptions following personality test feedback. These data indicate that in the future, precise appraisals of the impact of others' judgments on self-perceptions will require acknowledging the association between subject characteristics and the nature of the judgments given.

Significance of Manipulated Feedback

Having considered some factors that can affect the impact of others' ratings, we turn to an examination of the broader significance of feedback manipulation studies, particularly the degree of influence that feedback in such studies has been shown to have. Issues to be considered here are how long feedback effects last, their situational specificity, and the influence of feedback about a specific attribute for self-appraisal on other attributes.

One important but relatively neglected issue is the longevity of the impact of others' evaluations. Only two studies have examined

the effect of others' appraisals over time. In one investigation, subjects were given positive or negative feedback about physical skills by an expert, and their self-perceptions were reassessed immediately and after 1 day, 6 days, and 6 weeks (Haas & Maehr, 1965). Both positive and negative evaluations affected self-perceptions, and these effects were maintained over the 6-week period, although they appeared to diminish over time. Changes in dimensions not specifically evaluated were evident immediately after the evaluation, but were insignificant thereafter. Hicks (1962) gave subjects feedback that classmates judged them more favorably than their own self-perceptions on a group of personality traits. Two days after the initial evaluation, subjects were more likely to have raised their self-judgments on the elevated traits than on the control traits, although this difference did not hold after a week. Thus, the minimal evidence available on this issue suggests that the impact of others' judgments on self-perceptions holds over short periods of time but tends to diminish as time passes.

Also relevant in assessing the importance of feedback from others is the extent to which the effect of feedback generalizes from focal attributes to other characteristics. The three studies that have examined this effect used expert sources and systematically varied the relatedness of secondary attributes to the focal dimension (Haas & Maehr, 1965; Maehr et al., 1962; Videbeck, 1960). They found, not surprisingly, that judgments changed more on the dimension that was evaluated than on the one that was not (Maehr et al., 1962) and that those changes that did occur in other dimensions dissipated over time (Haas & Maehr, 1965). Therefore, relatively little information exists regarding the manner and extent to which content-focused evaluations are generalized to other characteristics of oneself.

Situational factors may also influence the degree of acceptance of others' self-evaluations, since the functional utility of accepting or rejecting others' impressions may vary from situation to situation. When college students feel that evaluations of their performance on a test are going to be made public, for example, they change their self-perceptions regarding that attitude more than do subjects who feel their responses will be known only to themselves (Gerard, 1961). Eagly and Acksen (1971) found that individuals changed their self-perceptions more in the direction of negative information and less in the direction of positive information when they felt that they would be retested on the attribute on which their performance was assessed, as compared with when they felt no retesting would occur. Positive attributes may be accepted and negative attributes may be fended off if there is no immediate prospect that the accuracy of these self-enhancing beliefs will be challenged. Other potential costs and gains of accepting or rejecting others' evaluations might also be envisioned. For instance, acknowledgment of certain positive attributes might be accompanied by the anticipation of favorable future outcomes or of increased demands from others. Similarly, the endorsement of negative attributes might lead to the anticipation of social rejection or loss of other favorable outcomes. In examining such problems it is important to distinguish between self-presentation and self-perception, since certain external factors might influence the manner in which people present themselves without affecting their actual self-perceptions.

The factors that most limit the interpretation of these manipulated feedback studies are the demand characteristics of the situation in which changes in self-perception are assessed. Invariably the appraisal of changes in self-evaluation was made in the presence of the evaluator or experimenter. When the evaluator is present, subjects who do not change their self-perceptions directly discredit the evaluator's appraisal, which may be difficult, particularly if the evaluator is presented as an expert. Even when evaluators are absent, experimenters may be perceived as being likely to communicate with them. Very rarely are there clearly reported efforts to disguise the postmanipulation self-appraisal process (Shrauger & Rosenberg, 1970).

One major way that the significance of manipulated feedback studies might be en-

hanced would involve making the assessment of change less reactive and more subtle. For example, the appraisal might be woven into some other aspect of the experiment supposedly unrelated to the portion in which feedback was given, as has been done in counterattitudinal advocacy studies (e.g., Rosenberg, 1965; Hendrick & Seyfried, 1974). Another possibility is to have the final self-evaluation made after an initial "debriefing," with the evaluation presented in the context of an appraisal of the effects of psychological experiments on individuals' attitudes and feelings.

A final issue in manipulated feedback studies is whether changes in self-evaluation are specific only to the self or can reflect modifications in judgments of others as well. There is evidence (Bramel, 1962; Edlow & Kiesler, 1966; Steiner, 1968) that when people are confronted with information discrepant from their self-evaluations, they not only change their self-evaluations but also modify their evaluations of others on the attribute judged. This may reflect a process of defensive projection or simply a change in the criteria they use for evaluating the attribute in question. Unfortunately most studies have looked only at shifts in the absolute level of self-judgments and not at changes in judgments of self relative to others. Such relative appraisals may be at least as significant as absolute judgments. Therefore, the effect of feedback on judgments of others as well as of oneself should be evaluated.

Discussion and Conclusions

The numerous studies of naturalistic and manipulated feedback that we have reviewed have had much to say about the relationship between others' judgments and self-appraisals; it is unfortunate that the flaws and limitations of these investigations have rendered the significance and validity of their findings questionable. Although there is evidence that individuals' self-perceptions and their views of others' perceptions of them are quite congruent, there is less evidence that self-perceptions are related to or influenced by others' actual perceptions. None of the studies of naturally occurring interactions were designed so that they would demonstrate unequivocally that receiving content-focused feedback from others leads to corresponding changes in one's own self-perceptions. In contrast, there is ample evidence of changes in self-perceptions following controlled feedback in laboratory settings. However, the importance of these findings is unclear because of the short-term nature of most assessments and the potential effects of demand characteristics. In evaluating the contributions and limitations of the available research, we give some attention to how information from others about the self is transmitted, received, interpreted, and acted upon. These are aspects of social self-perception that have for the most part been neglected by researchers in this area.

Availability of Evaluative Information

That there is minimal agreement between individuals' judgments of others' perceptions of them and their actual perceptions suggests that the communication of feedback to others may often be infrequent or ambiguous. Although norms regarding the evaluation of other people's behavior probably vary widely across different subcultures and situations, strong sanctions are often maintained against making direct appraisals, particularly when they are negative. In some of the only research on the communication of evaluations, Blumberg (1972) found that people report inhibiting the direct communication of all types of evaluations to others, particularly if it is negative or if the recipient is not known well. Barriers to direct expression can be found in intimate relationships as well as in more impersonal social interactions. This "not-even-your-best-friend-will-tell-you" phenomenon has been noted by Goffman (1955), who pointed out that unfavorable evaluations of close associates are typically given only when directly solicited and that in such a situation, chances are that the asker has already made some negative self-appraisal. Perhaps this accounts in part for the popularity of sensitivity training, in which people have the privilege of finding out what others

really think of them, and of assertiveness training, in which they can learn to communicate their true feelings about others.

To understand the real impact of others' opinions, one must determine how frequently such opinions are communicated directly in people's everyday social interactions. Who gives evaluations? On what dimensions? Under what circumstances? How often and how explicitly? The answers to such questions would facilitate an assessment of the relative influence of others' judgments on self-perceptions, as opposed to the opposite influence of self-perceptions on the perception of others' judgments. When information from others is not explicit, its interpretation may depend substantially on one's own self-perception on the attribute being assessed. In clinical contexts, for example, if people have concerns about what others think of them, it is frequently assumed that their inferences about others' feelings reflect a projection of their own self-evaluations.

It is quite likely that direct feedback occurs extensively in the socialization of young children by parents and other adults. During the process of language development, for instance, it seems certain that children come to model the construct system of those around them and to apply these constructs to themselves. Symbolic interactionists (Cooley, 1902; Mead, 1934) and self-perception theorists (Bem, 1967, 1972; Duval & Wicklund, 1972) alike have discussed the importance of preschool interactions in the development of a concept of self. It is surprising, however, that an empirical literature substantiating these arguments is nonexistent. Naturalistic studies of self-concept and perceived or actual assessments by others pick up developing selves as they enter the captive environment of elementary school. The subjects in these studies are typically in at least third or fourth grade (see Tables 1 and 2); only two studies have used first graders (Alberti, 1971; Trickett, 1969). Studies of controlled feedback almost exclusively use undergraduates. Since the preschool years are so vital to theories of the development of self-concept, it seems imperative that this period be attended to empirically. However, this may be easier to recommend than to implement. Trickett, for example, has noted the difficulties encountered in assessing the self-concept of first graders, which means that new and imaginative methods are necessary in this regard. Furthermore, recent work raises questions about whether young children possess the abstract concepts necessary to process information from others and use it in forming perceptions of themselves (Herzberger, Note 1). A naturalistic study of parent–child evaluative interactions might be a desirable first step in determining just what kind of feedback is given in the earliest stages of life.

Finally, in considering the availability of information from others it is important to recognize that people who are evaluated may help to determine how much evaluative information they receive. People's frequency of social interaction, how directly they ask for information, and how much they behave in ways that might elicit others' comments may all affect the amount of evaluative feedback received.

Interpretation of Information From Others

Although it is likely that people differ in their interpretations of others' feedback, particularly if that feedback is not explicit, these differences have not been explored extensively. People may disagree about what cues from others constitute an evaluation. And even when cues have been identified, people may differ in the inferences or conclusions they draw from these cues about others' judgments of them. For instance, it might be important to examine the extent to which information is considered principally for its specific content or for its evaluative meaning. To date the evidence suggests that content-specific feedback changes self-descriptions principally for those attributes on which feedback is given and only minimally on other attributes (Haas & Maehr, 1965; Videbeck, 1960). However, the nature of the situation in which these data were obtained may have maximized the impersonal, objective quality of evaluations and minimized the generalization that can occur in other contexts.

Characteristics of the evaluator may also be significant in determining the extent to which information is accepted. To date examinations of evaluator competence (Webster & Sobieszek, 1974) imply that only competence relevant to the attribute being judged has real impact on the acceptance of information. Expertise of the evaluator may be more complex, however, when the attributes judged do not involve specific, clearly defined skills. In these more subjective judgments, evaluators' competence may be judged more on global indices of status or on the extent to which they are perceived to hold norms similar to one's own on the dimensions in question.

A more situational aspect of the evaluator's competence involves whether or not the evaluator has a sufficient sample of one's behavior to make an adequate appraisal. Even if an appraiser is viewed as a good judge, his or her evaluation may be discounted if it is based on a limited or unrepresentative sample of behavior. Wyer, Henninger, and Wolfson (1975) showed, for example, that observers were much more likely to base their judgments on the limited behavior sample that they observed than were actors, whose self-appraisals were based less on that specific behavior sample and more on previous experiences.

Finally, the interpretation of the evaluation may depend on a perception of how candid other people are being. If one believes that there is some ulterior motive in making the evaluation (e.g., ingratiation or one-upmanship), it may not have as much effect on one's self-perception as a communication interpreted as more genuine.

Comparison With Self-Evaluations

An important aspect of others' judgments is how closely they agree with one's initial self-appraisal. Although judgments that match an initial self-perception may do little more than fortify this perception, judgments that are at variance frequently set up some dissonance or tension that requires cognitive reappraisal. There is an implicit disagreement between symbolic interactionist and self-attribution theories as to how such discrepancies are resolved. The symbolic interactionist view implies that such discrepancies are typically dealt with by changing one's self-perceptions, whereas self-attribution theories suggest that people have a reasonably clear and stable picture of themselves and may not readily conform to the discrepant appraisal of another individual.

The extent to which self-perception is maintained in the face of contradicting information from others presumably depends on the certainty of an individual's initial self-perceptions. Several factors may influence people's assuredness about their self-perceptions, all of which are related to opportunities for examining their own behavior. One factor is the salience of the dimension on which a judgment is made. Individuals are expected to have more clearly developed opinions about themselves on dimensions that are more important to them. A second aspect regarding the opportunity for observation may be the degree to which the person can compare his or her behavior with that of other people (cf. Festinger, 1954). Impressions may be more firmly established if people have the chance to compare themselves with other individuals. However, the opportunity for such comparisons may vary depending on the dimension being judged. A final determinant of assuredness may be the clarity of the criteria against which attributes are judged. A person is more likely to have a firmly established self-appraisal on an attribute that has a very clear public definition. One reason for children's potential susceptibility to self-concept molding may be their lack of clear criteria for defining particular characteristics. This may also account for the clinical observation that negative global self-perceptions (e.g., "I am rotten" or "I am a total failure") are resistant to change without exploration of what those attributes actually entail.

One complication in assessing the impact of others' feedback is that some changes in self-perception might be attributed to input from others when in fact they really reflect changes in individuals' independent appraisals of themselves. In the naturalistic studies

cited previously, changes toward others' perceptions could be accounted for by the individuals having changed or reappraised their own behavior. Certainly there is little in this literature that would negate the potential significance of the claim of self-perception theories that most self-knowledge comes from direct observation of one's own actions.

Maintenance of Changes

As previously mentioned, there is little evidence of the long-term effects of others' judgments on self-appraisals, and more adequate investigations of these effects are clearly required. Although these investigations would ideally involve naturally occurring interactions, manipulated feedback designs could also be employed. The use of negative feedback in such studies would, of course, be unacceptable ethically, but the effects of positive feedback could feasibly be investigated.

Long-term investigations are particularly important, since at least three processes may mitigate the impact of others' evaluations over time. First, discrepant feedback tends to be distorted so that it becomes more congruent with one's own initial self-perceptions (Harvey et al., 1957; Steiner, 1968; Suinn, Osborne, & Page, 1962). This tendency toward distortion has been demonstrated in experimental situations, although it is unclear how extensively such distortions occur in real-life settings.

A second mitigating factor may be that evaluations from another person may sometimes induce people to change their behavior in an opposite direction. If, for instance, an individual were evaluated as being self-centered but did not like that attribute, he or she might expend a special effort to be more altruistic and accordingly strengthen this perception of altruism. It has been shown that when subjects are told that they are making shorter or slower responses than those of other individuals, they lengthen and speed up their subsequent responses (Burnstein & Zajonc, 1965; Kleinke, 1975). Thus, the long-range impact of others' judgments may sometimes be to produce either no change in self-ratings or even changes in the opposite direction.

A third long-term effect of others' feedback may be that people change their social interactions so that they minimize their exposure to evaluators or to situations in which such feedback is likely to occur. Conceivably these mitigating long-term effects could be offset by an opposing tendency for people to change their behavior and also their self-perceptions to conform to others' role expectations. Unfortunately there are yet no investigations that have sorted out these potential outcomes.

Some Neglected Aspects of Others' Influence

It should be noted that empirical investigations of Mead and Cooley's looking-glass-self hypothesis have explored almost exclusively the impact of direct feedback from others. There may, however, be several less direct but equally important effects of others' judgments on self-perception. Simply being in the presence of others may influence the manner in which people behave (Goffman, 1959) and presumably come to evaluate their own behavior. At a conscious level one might deliberately enhance socially desirable and minimize socially undesirable behaviors when in the presence of others, and such changes could influence how one saw oneself. Less deliberately controlled aspects of behavior may also be affected by others' presence, as suggested in studies of audience effects on performance (e.g., Zajonc, 1965) and on self-evaluations of competence (Shrauger, 1972). Also, as Mead's (1934) notion of the generalized other implies, the physical presence of others is not imperative, so long as the perceiver can manage a mental impression of them.

Other individuals may also influence one's self-judgments by the manner in which they interact with people. Whether or not one receives help from a co-worker, for example, has been shown to affect one's subsequent self-esteen (Fisher & Nadler, 1974, 1976). In certain role relationships, such as that between a boss and subordinate, many interpersonal behaviors become quite clearly pre-

scribed. The nature of these interactions may convey to the individuals involved a certain degree of competence, or self-worth, without any explicit communication of these qualities over occurring. Although such processes have been described often in role theory (Goffman, 1955, 1959; Scheff, 1966), they have much less frequently been explored empirically, particularly with reference to their effects on people's self-perceptions.

A third indirect way that social interaction may influence self-perceptions is by affording the opportunity for people to compare their behavior with that of other people. Social comparison obviously requires the presence of other people at some point. It does not, however, prevent people from being active, reflective observers of their own behavior. The observance of others' behavior provides relative standards against which one's own actions and attributes may be judged. Although the significance of such comparison processes has long been recognized, few studies have explored how attributes of those against whom one compares oneself influence self-evaluation (Fontaine, 1974; Morse & Gergen, 1970; Strong & Gray, 1972). Morse and Gergen's investigation found that job applicants' judgments of themselves were substantially influenced by the apparent competence and appearance of other potential applicants. Perhaps even in situations that do not pull so explicitly for social comparisons, the observation of others' actions affects one's self-perceptions.

Finally, other people may indirectly affect one's self-perceptions when they are observed making evaluations of other individuals. Even if people do not receive feedback directly, observing someone make a judgment of another individual may provide indirect information about how they are viewed by the evaluator. How much this actually occurs depends of course on how explicit the criteria for evaluating the other person's behavior are and on the degree to which one sees similarity or dissimilarity between one's own behavior and that of the person being evaluated.

In sum, it may be that the aspect of the looking-glass-self hypothesis that has been most frequently examined, the effect of direct feedback from other people, reflects only one of the ways that interaction with others has an impact on self-judgments. Furthermore, this means of influence may well be of no greater importance than the others. The relative ease with which direct evaluation can be explored ought not to preclude the examination of other viable aspects of social interaction that may also lead to the modification of self-evaluations.

Reference Note

1. Herzberger, S. The development of social self-perception. In J. Chevalier & R. Wolfe (Chairs), *Through a lens clearly: Enhancing a dim view of person perception*. Symposium presented at the State University of New York College at Geneseo, October 1978.

References

Alberti, J. M. Self-perception-in-school: Validation of an instrument and a study of the structure of children's self-perception-in-school, and its relationship to school achievement, behavior and popularity (Doctoral dissertation, State University of New York at Buffalo, 1970). *Dissertation Abstracts International*, 1971, *31*, 4535A–4536A. (University Microfilms No. 71-6048)

Amatora, M. Validity in self-evaluation. *Educational and Psychological Measurement*, 1956, *16*, 119–126.

Ausubel, D. P. Socioempathy as a function of sociometric status in an adolescent group. *Human Relations*, 1955, *8*, 75–84.

Ausubel, D. P., & Schiff, H. M. Some intrapersonal and interpersonal determinants of individual differences in socioempathic ability among adolescents. *Journal of Social Psychology*, 1955, *41*, 39–56.

Ausubel, D. P., Schiff, H. M., & Gasser, E. B. A preliminary study of developmental trends in socioempathy: Accuracy of perception of own and others' sociometric status. *Child Development*, 1952, *23*, 111–128.

Backman, D., Secord, P., & Pierce, J. Resistance to change in the self-concept as a function of consensus among significant others. *Sociometry*, 1963, *26*, 102–111.

Baldwin, J. M. *Social and ethical interpretations in mental development: A study in social psychology*. New York: Macmillan, 1897.

Bem, D. J. Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 1967, *74*, 183–200.

Bem, D. J. Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psy-*

chology (Vol. 6). New York: Academic Press, 1972.

Bergin, A. E. The effect of dissonant persuasive communications upon changes in a self-referring attitude. *Journal of Personality,* 1962, *30,* 423–438.

Binderman, R. M., Fretz, B. R., Scott, N. A., & Abrams, M. H. Effects of interpreter credibility and discrepancy level of results on responses to test results. *Journal of Counseling Psychology,* 1972, *19,* 399–403.

Bishop, J. B. Another look at counselor, client and supervisor ratings of counselor effectiveness. *Counselor Education and Supervision,* 1971, *10,* 319–323.

Bledsoe, J. C., & Wiggins, R. G. Congruence of adolescents' self-concepts and parents' perceptions of adolescents' self-concepts. *Journal of Psychology,* 1973, *83,* 131–136.

Blumberg, H. H. Communication of interpersonal evaluations. *Journal of Personality and Social Psychology,* 1972, *23,* 157–162.

Bramel, D. A. A dissonance theory approach to defensive projection. *Journal of Abnormal and Social Psychology,* 1962, *69,* 121–129.

Brams, J. Counselor characteristics and effective communication in counseling. *Journal of Counseling Psychology,* 1961, *8,* 25–30.

Breslin, H. B. The relationship between the physically handicapped child's self-concept and his peer reputation (Doctoral dissertation, Oregon State University, 1968). *Dissertation Abstracts International,* 1968, *29,* 1493B. (University Microfilms No. 68-14,871)

Buckley, E. F. The relationship between student-teacher perceptions and pupil perceptions of the student-teacher (Doctoral dissertation, North Texas State University, 1970). *Dissertation Abstracts International,* 1970, *30,* 4672A–4673A. (University Microfilms No. 70-9121)

Burke, P. J. Some preliminary data on the use of self-evaluations and peer ratings in assigning university grades. *Journal of Educational Research,* 1969, *62,* 444–448.

Burns, R. To a louse. In, *Poems*. New York: Heritage Press, 1965.

Burnstein, E., & Zajonc, R. B. Individual task performance in a changing social structure. *Sociometry,* 1965, *28,* 16–29.

Carroll, J. B. Ratings on traits measured by a factored personality inventory. *Journal of Abnormal and Social Psychology,* 1952, *47,* 626–632.

Cogan, L. C., Conklin, A. M., & Hollingworth, H. L. An experimental study of self-analysis, estimates of associates, and the results of tests. *School and Society,* 1915, *2,* 171–179.

Cooley, C. H. *Human nature and the social order.* New York: Scribner's, 1902.

Cooper, J., & Duncan, B. L. Cognitive dissonance as a function of self-esteem and logical inconsistency. *Journal of Personality,* 1971, *39,* 289–302.

Davidson, H. H., & Lang, G. Children's perceptions of their teachers' feelings toward them related to self-perception, school adjustment and behavior. *Journal of Experimental Education,* 1960, *29,* 107–118.

De Jung, J. E., & Gardner, E. F. The accuracy of self-role perceptions: A developmental study. *Journal of Experimental Education,* 1962, *31,* 27–41.

Denzin, N. K. The significant others of a college population. *Sociological Quarterly,* 1966, *7,* 298–310.

Douce, P. D. M. Selected aspects of personality related to social acceptance and clothing-oriented variables (Doctoral dissertation, Utah State University, 1969). *Dissertation Abstracts International,* 1970, *30,* 3730B. (University Microfilms No. 70-2392)

Duval, S., & Wicklund, R. A. *A theory of objective self-awareness.* New York: Academic Press, 1972.

Dmitruk, V. M., Collins, R. W., & Clinger, D. L. The "Barnum effect" and acceptance of negative personal evaluation. *Journal of Consulting and Clinical Psychology,* 1973, *41,* 192–194.

Eagly, A. H. Involvement as a determinant of response to favorable and unfavorable information. *Journal of Personality and Social Psychology Monograph,* 1967, *7*(3, Pt. 2).

Eagly, A. H., & Acksen, B. A. The effect of expecting to be evaluated on change toward favorable and unfavorable information about oneself. *Sociometry,* 1971, *34,* 411–422.

Eagly, A. H., & Whitehead, G. I. Effect of choice on receptivity to favorable and unfavorable evaluations of oneself. *Journal of Personality and Social Psychology,* 1972, *22,* 223–230.

Edlow, D. W., & Kiesler, C. A. Ease of denial and defensive projection. *Journal of Experimental Social Psychology,* 1966, *2,* 56–69.

Eisenmann, R., & Robinson, N. Peer-, self- and test-ratings of creativity. *Psychological Reports,* 1968, *23,* 471–474.

Erikson, E. H. The problem of ego identity. *Journal of the American Psychoanalytic Association,* 1956, *4,* 56–121.

Evans, G. C. The influence of "fake" personality evaluations on self-descriptions. *Journal of Psychology,* 1962, *53,* 457–463.

Festinger, L. A theory of social comparison processes. *Human Relations,* 1954, *7,* 117–140.

Fey, W. F. Acceptance by others and its relation to acceptance of self and others: A revaluation. *Journal of Abnormal and Social Psychology,* 1955, *50,* 274–276.

Fisher, J. D., & Nadler, A. The effect of similarity between donor and recipient on recipient's reactions to aid. *Journal of Applied Social Psychology,* 1974, *4,* 230–243.

Fisher, J. D., & Nadler, A. Effect of donor resources on recipient self-esteem and self-help. *Journal of Experimental Social Psychology,* 1976, *12,* 139–150.

Fontaine, G. Social comparison and some determinants of expected personal control and expected performance in a novel task situation. *Journal of Personality and Social Psychology*, 1974, *29*, 487–496.

Friedsam, H. J., & Martin, H. W. A comparison of self and physicians' health ratings in an older population. *Journal of Health and Human Behavior*, 1963, *4*, 179–183.

Gerard, H. B. Some determinants of self-evaluation. *Journal of Abnormal and Social Psychology*, 1961, *62*, 288–293.

Goffman, E. On face-work: An analysis of ritual elements in social interaction. *Psychiatry: Journal for the Study of Interpersonal Processes*, 1955, *18*, 213–231.

Goffman, E. *The presentation of self in everyday life*. New York: Anchor Books, 1959.

Goldings, H. J. On the avowal and projection of happiness. *Journal of Personality*, 1954, *23*, 30–47.

Goodman, S. A. A further exploration of the relationship between self-concept and sociometric status (Doctoral dissertation, University of North Carolina at Chapel Hill, 1973). *Dissertation Abstracts International*, 1973, *34*, 170A. (University Microfilms No. 73-16,476)

Gordon, C., & Gergen, K. J. (Eds.). *The self in social interaction* (Vol. 1). New York: Wiley, 1968.

Goslin, P. A. Accuracy of self-perception and social acceptance. *Sociometry*, 1962, *25*, 283–296.

Gray, D. F., & Gaier, E. L. The congruency of adolescent self-perceptions with those of parents and best friends. *Adolescence*, 1974, *9*, 299–304.

Green, G. H. Insight and group adjustment. *Journal of Abnormal and Social Psychology*, 1948, *43*, 49–61.

Haas, H. I., & Maehr, M. L. Two experiments on the concept of self and the reaction of others. *Journal of Personality and Social Psychology*, 1965, *1*, 100–105.

Halperin, K., Snyder, C. R., Shenkel, R. J., & Houston, B. K. Effects of source status and message favorability on acceptance of personality feedback. *Journal of Applied Psychology*, in press.

Hamilton, D. L. Measures of self-esteem, dominance, and dogmatism: Convergent and discriminant validity. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 1969, *4*, 127–128. (Summary)

Hartup, W. W. Peer interactions and social organization. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology*. New York: Wiley, 1970.

Harvey, O. J., & Clapp, W. F. Hope, expectancy, and reactions to the unexpected. *Journal of Personality and Social Psychology*, 1965, *2*, 45–52.

Harvey, O. J., Hunt, D. E., & Schroder, H. M. *Conceptual systems and personality organization*. New York: Wiley, 1961.

Harvey, O. J., Kelley, H. H., & Shapiro, M. M. Reactions to unfavorable evaluations of the self made by other persons. *Journal of Personality*, 1957, *25*, 393–411.

Hase, H. D., & Goldberg, L. R. Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, 1967, *67*, 231–248.

Helper, M. M. Parental evaluations of children and children's self-evaluations. *Journal of Abnormal and Social Psychology*, 1958, *56*, 190–194.

Hendrick, C., & Seyfried, B. A. Assessing the validity of laboratory-produced attitude change. *Journal of Personality and Social Behavior*, 1974, *29*, 865–870.

Hicks, J. The influence of group flattery upon self-evaluation. *Journal of Social Psychology*, 1962, *58*, 147–151.

Horowitz, F. D. The relationship of anxiety, self-concept, and sociometric status among fourth, fifth, and sixth grade children. *Journal of Abnormal and Social Psychology*, 1962, *65*, 212–214.

Israel, J. Self-evaluation in groups. *Acta Sociologica*, 1958, *3*, 29–47.

James, W. *The principles of psychology*. New York: Holt, 1890.

Jansen, D. G., Robb, G. P., & Bonk, E. C. Peer ratings and self-ratings on twelve bipolar items of practicum counselors ranked high and low in competence by their peers. *Journal of Counseling Psychology*, 1973, *20*, 419–424.

Jones, E. E., Gergen, K. J., & Davis, K. Some determinants of reactions to being approved or disapproved as a person. *Psychological Monographs*, 1962, *76*(2, Whole No. 521).

Jones, E. E., & Nisbett, R. E. *The actor and the observer: Divergent perceptions of the causes of behavior*. Morristown, N.J.: General Learning Press, 1971.

Johnson, H. H. Some effects of discrepancy level on responses to negative information about one's self. *Sociometry*, 1966, *29*, 52–66.

Jorgenson, G. Q. Relationships among self, peer and teacher evaluations of motivational dispositions and behaviors in third grade (Doctoral dissertation, University of Utah, 1967). *Dissertation Abstracts*, 1967, *27*, 3314A–3315A. (University Microfilms No. 67-3132)

Jourard, S. M., & Remy, R. M. Perceived parental attitudes, the self, and security. *Journal of Consulting Psychology*, 1955, *19*, 364–366.

Kelman, H. C., & Parloff, M. B. Interrelations among three criteria of improvement in group therapy: Comfort, effectiveness, and self-awareness. *Journal of Abnormal and Social Psychology*, 1957, *54*, 281–288.

Kemper, T. D. Self-conceptions and the expectations of significant others. *Sociological Quarterly*, 1966, *7*, 323–343.

Kinch, J. W. A formalized theory of the self-concept. *American Journal of Sociology*, 1963, *68*, 481–486.

Kinch, J. W. Experiments on factors related to self-concept change. *Journal of Social Psychology*, 1968, *74*, 251–258.

Kleinke, C. L. Effects of false feedback about response length on subjects' perception of an interview. *Journal of Social Psychology*, 1975, *95*, 99–104.

Klimoski, P. J., & London, M. J. Role of the rater in performance appraisal. *Journal of Applied Psychology*, 1974, *59*, 445–451.

Lomont, J. F. Repressors and sensitizers as described by themselves and their peers. *Journal of Personality*, 1966, *34*, 224–240.

Maehr, M. L., Mensing, J., & Nafzger, S. Concept of self and the reactions of others. *Sociometry*, 1962, *25*, 353–357.

Manis, M. Social interaction and the self-concept. *Journal of Abnormal and Social Psychology*, 1955, *51*, 362–370.

Marcia, J. Ego-identity status: Relationship to change in self-esteem, "general maladjustment" and authoritarianism. *Journal of Personality*, 1967, *35*, 119–133.

Mayo, G. D., & Manning, W. H. Motivation measurement. *Educational and Psychological Measurement*, 1961, *21*, 73–83.

McConnell, G. Questions about you. *Education*, 1959, *80*, 112–114.

McIntyre, C. J. Acceptance by others and its relation to acceptance of self and others. *Journal of Abnormal and Social Psychology*, 1952, *47*, 624–625.

Mead, G. H. *Mind, self and society*. Chicago: University of Chicago Press, 1934.

Meehl, P. E. Wanted—A good cookbook. *American Psychologist*, 1956, *11*, 263–272.

Miyamoto, S. F., & Dornbusch, S. A test of the symbolic interactionist hypotheses of self-conception. *American Journal of Sociology*, 1956, *61*, 399–403.

Morse, S., & Gergen, K. J. Social comparison, self-consistency, and the concept of self. *Journal of Personality and Social Psychology*, 1970, *16*, 148–156.

Mosher, D. L. Approval motive and acceptance of "fake" personality test interpretations which differ in favorability. *Psychological Reports*, 1965, *17*, 395–402.

Mote, F. B. The relationship between child self-concept in school and parental attitudes and behaviors in child rearing (Doctoral dissertation, Stanford University, 1966). *Dissertation Abstracts*, 1967, *27*, 3319A. (University Microfilms No. 67-4309)

Newcomb, T. M. *The acquaintance process*. New York: Holt, Rinehart & Winston, 1961.

Orpen, C., & Bush, R. The lack of congruence between self-concept and public image. *Journal of Social Psychology*, 1974, *93*, 145–146.

Papageorgis, D., & McCann, B. M. Effect of discrepant communications on self-perception. *Journal of Social Psychology*, 1965, *67*, 227–237.

Peirce, C. S. Questions concerning certain faculties claimed for man. *Journal of Speculative Philosophy*, 1868, *2*, 103–114.

Perkins, H. V. Teachers' and peers' perception of children's self-concepts. *Child Development*, 1958, *29*, 203–220.

Peterson, D. R. Scope and generality of verbally defined personality factors. *Psychological Review*, 1965, *72*, 48–59.

Phillips, B. N. Age changes in accuracy of self-perception. *Child Development*, 1963, *64*, 1041–1046.

Powell, M. G. Comparison of self-rating, peer-ratings, and experts'-ratings of personality adjustment. *Educational and Psychological Measurement*, 1948, *8*, 225–234.

Quarantelli, E. L., & Cooper, J. Self-conceptions and others: A further test of the Meadian hypothesis. *Sociological Quarterly*, 1966, *7*, 281–297.

Raven, B. H., & Rubin, J. Z. *Social psychology: People in groups*. New York: Wiley, 1976.

Reeder, L. G., Donohue, G. A., & Biblarz, A. Conceptions of self and others. *American Journal of Sociology*, 1960, *66*, 153–159.

Reese, H. W. Relationship between self-acceptance and sociometric choice. *Journal of Abnormal and Social Psychology*, 1961, *62*, 472–474.

Regan, J. W., Gosselink, H., Hubsch, J., & Ulsh, E. Do people have inflated views of their own ability? *Journal of Personality and Social Psychology*, 1975, *31*, 295–301.

Rokeach, M. Studies in beauty: II. Some determiners of the perception of beauty in women. *Journal of Social Psychology*, 1945, *22*, 155–169.

Rosenberg, M. J. When dissonance fails: On eliminating evaluation apprehension from attitude measurement. *Journal of Personality and Social Psychology*, 1965, *1*, 28–42.

Rosengren, W. R. The self in the emotionally disturbed. *American Journal of Sociology*, 1961, *66*, 454–462.

Scheff, T. J. *Being mentally ill: A sociological theory*. Chicago: Aldine, 1966.

Schneider, B. Relationship between various criteria of leadership in small groups. *Journal of Social Psychology*, 1970, *82*, 253–261.

Schneider, D. J., *Social psychology*. Reading, Mass.: Addison-Wesley, 1976.

Scott, W. A., & Johnson, R. C. Comparative validity of direct and indirect personality tests. *Journal of Consulting and Clinical Psychology*, 1972, *38*, 301–318.

Secord, P. F., & Backman, C. W. *Social psychology*. New York: McGraw-Hill, 1974.

Sherwood, J. J. Self-identity and referent others. *Sociometry*, 1965, *28*, 66–81.

Sherwood, J. J. Increased self-evaluation as a function of ambiguous evaluations by referent others. *Sociometry*, 1967, *30*, 404–409.

Shrauger, J. S. Self-esteem and relations to being observed by others. *Journal of Personality and Social Psychology*, 1972, *23*, 192–200.

Shrauger, J. S., & Lund, A. Self-evaluation and reactions to evaluations from others. *Journal of Personality*, 1975, *43*, 94–108.

Shrauger, J. S., & Rosenberg, S. E. Self-esteem and the effects of success and failure feedback on performance. *Journal of Personality*, 1970, *38*, 404–417.

Snyder, C. R., & Shenkel, R. J. Effects of "favorability," modality, and relevance on acceptance of general personality interpretations prior to and after receiving diagnostic feedback. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 34–41.

Sobieszek, B. I., & Webster, M. Conflicting sources of evaluations. *Sociometry*, 1973, *36*, 550–560.

Steiner, I. D. Reactions to adverse and favorable evaluations of one's self. *Journal of Personality*, 1968, *36*, 553–563.

Strong, S. R., & Gray, B. L. Social comparison, self-evaluation, and influence in counseling. *Journal of Counseling Psychology*, 1972, *19*, 178–183.

Suinn, R. M., Osborne, D., & Page, W. The self-concept and accuracy of recall of inconsistent self-related information. *Journal of Clinical Psychology*, 1962, *18*, 473–474.

Sundberg, N. D. The acceptability of "fake" versus "bona fide" personality test interpretations. *Journal of Abnormal and Social Psychology*, 1955, *50*, 145–147.

Swanson, B. M. Parent-child relations: A child's acceptance by others, of others and of self (Doctoral dissertation, University of Oklahoma, 1969). *Dissertations Abstracts International*, 1969, *30*, 1890B. (University Microfilms No. 69-16,576)

Tagiuri, R., Blake, R. R., & Bruner, J. S. Some determinants of the perception of positive and negative feelings in others. *Journal of Abnormal and Social Psychology*, 1953, *48*, 585–592.

Teichman, M. Cognitive differentiation between self-concept and image of self ascribed to parents in boys on the verge of delinquency. *Perceptual and Motor Skills*, 1972, *34*, 573–574.

Todorosky, N. R. Self-acceptance and acceptance by peers (Doctoral dissertation, Michigan State University, 1972). *Dissertation Abstracts International*, 1972, *33*, 2359B–2360B. (University Microfilms No. 72-30,054)

Trickett, H. V. Stability and predictability of children's self-concept and perception by others: A developmental study (Doctoral dissertation, Stanford University, 1968). *Dissertation Abstracts International*, 1969, *29*, 2577A. (University Microfilms No. 69-306)

Tschechtelin, S. M. A. Self-appraisal of children. *Journal of Educational Research*, 1945, *39*, 25–32.

Videbeck, R. Self-conception and the reactions of others. *Sociometry*, 1960, *23*, 351–359.

Walhood, D. S., & Klopfer, W. G. Congruence between self-concept and public image. *Journal of Consulting and Clinical Psychology*, 1971, *37*, 148–150.

Webb, W. B. Self-evaluations, group evaluations, and objective measures. *Journal of Consulting Psychology*, 1955, *19*, 210–212.

Webster, M., & Sobieszek, B. I. *Sources of self-evaluation: A formal theory of significant others and social influence*. New York: Wiley, 1974.

Weisberg, P. Student acceptance of bogus personality interpretations differing in level of social desirability. *Psychological Reports*, 1970, *27*, 743–746.

Werdelin, I. A study of the relationship between teacher ratings, peer ratings and self ratings of behavior in school. *Scandinavian Journal of Educational Research*, 1969, *13*, 147–169.

Winthrop, H. Self-images of personal adjustment vs. the estimates of friends. *Journal of Social Psychology*, 1959, *50*, 87–99.

Wyer, R. S., Henninger, M., & Wolfson, M. Informational determinants of females' self-attributions and observers' judgments of them in an achievement situation. *Journal of Personality and Social Psychology*, 1975, *32*, 556–570.

Wylie, R. C. *The self-concept* (Vol. 1, Rev. ed.). Lincoln: University of Nebraska Press, 1974.

Zajonc, R. B. Social facilitation. *Science*, 1965, *149*, 269–274.

Ziller, R. C. *The social self*. New York: Pergamon Press, 1973.

Received January 24, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 3, 574–595

Sex Differences in Childhood Psychopathology: A Review

Robert F. Eme
North Central College

Sex differences in the major categories of childhood behavior disorders most relevant to the issue of continuity between child and adult disorders are reviewed. Explanations for these differences are explored with attention given to both the different experiences and the different endowments of the sexes. These differences are then compared and contrasted with sex differences in adult psychopathology.

Although the sex differences in disturbed adults, both treated and untreated, have been extensively examined (Chesler, 1972; Dohrenwend & Dohrenwend, 1969; Garai, 1970; Gove & Tudor, 1973), sex differences in children appear to have been accorded only one such examination (Gove & Herb, 1974). Furthermore, this review is beset by the same limitation as part of a similar review of the adult literature by Gove and Tudor (1973), namely, as Dohrenwend and Dohrenwend (1975) suggested, the authors relied almost entirely on role theory to explain the sex differences, arguing that at some time or place one or the other sex is under greater stress and hence more prone to psychiatric disorder. It should also be noted that the Gove and Herb review was without the benefit of Maccoby and Jacklin's (1974) encyclopedic coverage of the psychology of sex differences and its subsequent critique by Block (1976, 1978). Hence, it is the purpose of the present review to extend Gove and Herb's earlier review in the hope of remediating its limitation.

In the absence of a generally accepted taxonomy (Achenbach & Edelbrock, 1978), the categories of child behavior disorders to be reviewed are those that have a special relevance to the issue of continuity between child and adult psychopathology (Kohlberg, LaCrosse, & Ricks, 1972; Rutter, 1972b). Childhood is considered as that period of the life span from birth to adolescence. The reason for this demarcation is that since adolescent sex differences in psychopathology closely resemble adult sex differences (Gove & Herb, 1974; Graham & Rutter, 1977; Rutter, 1974), it is the period prior to adolescence that requires attention. The findings, in accordance with Dohrenwend and Dohrenwend's suggestion, are analyzed not only in terms of the different experiences of the sexes but also in terms of the sexes' different endowments. These findings are then compared with those of the adult literature.

Finally, it should be noted that although expository convenience dictates a consideration of factors influencing sex differences under the separate rubrics of endowment or experience, in no way does the author construe these influences as polarized. Sex differences are construed here, as they are by virtually all theorists, to be multidetermined products of an interaction between biological and learning processes. This interaction has been variously considered by behavioral geneticists in terms of *norm of reaction* (McClearn & DeFries, 1973), by psychopathologists in terms of *diathesis–stress* (Rosenthal, 1970), by developmentalists in terms of a *greater readiness to learn* (Maccoby & Jacklin, 1974) and by psychosexual developmentalists in terms of an *imprimatur* (Money & Ehrhardt, 1972) that is shaped by a *social script* (Gagnon & Simon, 1973). The aptness of each

---

I wish to thank William Herr and Leslie Wilcoxson for their help in the preparation of this article.

Requests for reprints should be sent to Robert F. Eme, 1200 West Sherwin, Chicago, Illinois 60626.

metaphor varies with the idiomatic conventions of one's discipline, and it is left to the reader to decide which if any of the images is the most suitable.

Methodological Problems

The methodological problems in making a diagnosis of psychological disorder in epidemiological studies of adult populations have been cogently detailed by Dohrenwend and Dohrenwend (1974). These same problems obviously exist in similar studies of child populations, as Conger and Cole (1975) noted in a recent example. Although it is beyond the scope of this review to engage in a detailed analysis of these problems, there is a major difficulty that must be considered.

In their review of the adult literature Dohrenwend and Dohrenwend scored the inadequacy of studies based on treated rates. Hence their review focused for the most part on studies of untreated cases of psychopathology, of which they uncovered 70 done since the turn of the century. This same problem of focusing on treated rates exists in the studies of child psychopathology, but unfortunately the paucity of studies of untreated cases renders similar recourse impossible. For example, Gove and Herb (1974) relied exclusively on treated cases, while noting that scattered results for studies of non-clinic-attending populations were generally consistent with their formulations.

Although the scarcity of prevalence studies of untreated psychological disorders in children makes any conclusions about sex differences somewhat tenuous, it does not render such an attempt futile. Rather, following the lead of Gove and Herb, such studies can be employed in a supportive fashion to confirm or disconfirm findings based on treated cases. Furthermore, there is the comforting finding of past researchers that prevalence rates based on untreated cases generally confirm the sex differences established by the data based on treated cases (Anthony, 1970; Gove & Herb, 1974). Hence, although the data base used to investigate the existence of sex differences in child psychological disorders is limited in comparison with that of the adult literature, it seems sufficiently robust to warrant such an examination.

Adjustment Reaction of Childhood

The most common child diagnosis is that of adjustment reaction of childhood (Anthony, 1970). The determination of the diagnosis can include virtually any symptom or set of symptoms that appear to be precipitated by acute situational stress.

In this regard it is interesting to note the general agreement among reviewers (Kanner, 1960; Kohlberg et al., 1972; Rutter, 1972b) that there is little continuity between these characteristics and adult psychopathology. Hence, the most common interpretation given to these symptoms is that they represent manifestations of developmental stress occurring in essentially well-adjusted children.

Studies of both treated and untreated cases of adjustment reaction of childhood indicate greater prevalence among males than among females (Anthony, 1970; Gove & Herb, 1974). However, this greater prevalence does not manifest itself until the school-age years (Richman, Stevenson, & Graham, 1975). The typical explanation given for this disparity is the greater stress on males due to their biological immaturity, their temperament, expectations for their behavior, and the feminine environment in which they live (Bardwick, 1971; Gove & Herb, 1974). Thus Bardwick maintained that cultural pressure on girls is much less because their general predispositions and the cultural interpretations of what is acceptable are more nearly matched. This lesser pressure enables a girl to experience a less stormy childhood. Examples of greater cultural pressure cited by Bardwick are the greater expectations for the male to achieve; the greater censure of the male who is passive, withdrawn, and overdependent; and the message given the male to be active and aggressive and, at the same time, the threat of punishment when these actions result in nonconformity. Gove and Herb added that this greater cultural pressure impacts on a male who is less intellectually and physically mature and more

aggressive, thus engendering even greater stress.

It should be noted that at the time these hypotheses were offered, research substantiation was minimal. More recently, however, empirical support for some of them has been accumulating. For example, Rutter (1977b) concluded his review of the literature on temperament by stating that there is good evidence individual differences in this variable play an important role in the development of psychological problems. He scored the following type of child as being especially at risk: a child who is emotionally tense, who is slow to adapt to new situations, whose behavior is difficult to change, who has irregular eating, sleeping, and bowel habits, who tends to be irritable and negative in mood, and who is unusually tolerant of messiness and disorder. Though Rutter himself did not mention sex differences, Moss (1974) summarized 10 years of his work with infants, consisting of several independent studies, and concluded that males were generally more irritable than females. This conclusion was confirmed in a recent study by Phillips, King, and DuBois (1978) on newborns that controlled for the methodological limitations of prior studies, not the least of which was the variable of male circumcision. Thus it seems that at least one of the temperamental attributes listed by Rutter shows a sex difference.

Another example is in the area of differential pressure for achievement. In a recent review Hoffman (1977) indicated that although boys and girls are both encouraged to do well in school, some important sex differences in achievement pressures may exist. She cited observational and interview studies in which parental interactions with sons emphasized greater achievement and competition than did interactions with daughters. Parents are also more likely to want their sons to be hardworking and ambitious and are more disappointed if they do not achieve their academic and occupational goals.

Other research pertinent to the traditional explanations are taken up later. At this juncture, however, there are two other explanations that have generally been overlooked in the literature but that merit equal consideration in the delineation of the divergent endowments and experiences of the sexes.

The first is simply that the annoyance threshold for male deviance is less than that for female deviance. For example, Shepard, Oppenheim, and Mitchell (1966), in a study that compared 50 referrals with matching nonreferral controls, concluded that it was primarily parental reaction, not severity of disturbance, that dictated a clinical referral. It appeared that clinic-attending mothers felt more puzzled and hopeless in coping with their children's problems than did control mothers. They further reported that the most obvious difference lay in the number of nonclinic mothers who accepted their children's behavior as a temporary difficulty and that this increased tolerance was particularly obvious among the mothers of girls. Similarly, Chess and Thomas (1972), in their study of temperamental differences in children, indicated that parents were less tolerant of a lack of persistence and of distractibility in males than in females. Battle and Lacey (1972), in a study of hyperactivity in 74 subjects drawn from the Fels longitudinal study, reported that mothers of highly active males were critical, disapproving, unaffectionate, and severe in their punishment. No correlation of these maternal behaviors with high activity levels in females was found. Serbin and O'Leary (1975) observed teacher-child interaction in 15 preschool classrooms. They reported that a disruption by a male was more likely to elicit a reprimand than a similar disruption by a female; males' reprimands were also more severe. For example, teachers responded over three times as often to males who hit or broke things, and the boys usually got a loud public reprimand.

Hence, it may be that females—mothers and teachers—are more likely to view the same disturbance as more pathological in the male than in the female. And since they are the primary sources of referral and evaluation in epidemiological studies, there is the resultant excess of males with adjustment problems. This may be partly a function of the fact that adults feel less comfortable and competent with children of the opposite

sex (Sullivan, 1953). Hence it is not surprising that in a national survey, Hoffman (1977) indicated that some of the most common reasons women gave for preferring a female were "that girls are easier to raise and more obedient . . . are cuter, sweeter, or not as mean" (p. 648).

This explanation, though, should be tempered by the caveat that one is assuming the disturbances are indeed the same. Although this assumption seems reasonable in the studies of Chess and Thomas, Battle and Lacey, and Serbin and O'Leary, since the authors themselves clearly designated the behaviors as similar, it would be rash to extrapolate these findings of similarity to the multitude of adjustment problems that the sexes present. For example, both Garai and Scheinfeld (1968) and Feshbach (1970) indicated that aggression in girls is more likely to assume a prosocial mode, for example, rule enforcement, whereas male aggression is more likely to be destructive. Hence the "same" aggression is less tolerated in males because it is actually dissimilar in mode. Though Maccoby and Jacklin (1974) challenged the validity of this distinction, it still serves the function of illustrating the slipperiness of designating similarity.

A second reason for the greater prevalence of male adjustment reactions of childhood relates to a difference in endowment, namely, that males may be constitutionally more vulnerable not only to biological but to psychological stress as well. As the reviews by Garai and Scheinfeld, Maccoby and Jacklin, and Birns (1976) indicate, there is little doubt that despite the fact that males are larger and stronger than females at almost every age, they are more vulnerable to almost any kind of physical hazard; and this vulnerability is magnified by the effect of poverty (Birns, 1976). Though the ratio of male to female conceptions is 130:100, the ratio is reduced to 105:100 at birth in the United States. They suffer more abortions, stillbirths, miscarriages, prematurity, anoxia, and other birth complications. They are also more likely than females to suffer serious defects as a result of prematurity (Braine, Heimer, Wortis, & Freedman, 1966) or anoxia (Gottfried, 1973). During infancy 37% more males die, and throughout life males are more afflicted by the major diseases (Garai & Scheinfeld, 1968). They are also more likely to suffer ill effects from malnutrition (Tanner, 1970) and radiation (Rutter, 1972b).

The reasons commonly cited for this greater male vulnerability are greater male immaturity (Rutter, 1972b), greater male susceptibility to sex-linked diseases (Garai & Scheinfeld, 1968), and possible adverse maternal immunological reaction to male fetal tissue because of the presence of the male Y chromosome (Mussen, Conger, & Kagan, 1974). Analogously, there appears to be a greater male vulnerability to psychological stress.

Rutter (1970) conducted a study of 200 families and matching controls in which one of the parents was a psychiatric patient. The sample contained an equal number of male and female psychiatric patients who had one or more children under the age of 15. Information about the families was obtained by psychiatric interview of the parents, and information about the children's psychiatric state was obtained by teacher questionnaire. He found that discord and disruption in the home was consistently and strongly associated with anitsocial disorder in boys but not in girls. No consistent associations were found between family characteristics and neurosis in either boys or girls. Nor did the sex of the ill parent bear a relation to the likelihood of the children's developing a psychiatric disorder. He then in some detail considered and dismissed the following possible methodological biases as alternate explanations to this provocative sex difference in preadolescent children: The sex difference is peculiar to the children's behavior at school; teachers cannot perceive deviance in girls as well as in boys; and the sex difference is due to diagnostic differences. By reviewing other relevant studies he concluded and found that though the evidence is surprisingly meagre and to some extent contradictory when considered in relation to his clear-cut findings, it indeed appears that males are more vulner-

able to adverse effects of family discord and disruption.

Similarly, Wolkind and Rutter (1973) in their Isle of Wight survey reported that there was a strong tendency for children who were in short-term (6 months or less) care because of maternal confinement or physical illness to have a behavioral disturbance, as defined by teacher and parent interviews. The tendency was much more marked for boys than for girls. With long-term care, howevers, girls seem to be as susceptible to ill effects as are boys. Rutter (1972a), in his review of the literature on maternal deprivation, pointed out that the findings on sex differences are somewhat contradictory and that no differences have been found in many studies. However, where there has been a difference, the male has usually been found to be more vulnerable to the adverse effects of separation. Bowlby (1973) cited a finding parallel to Rutter's in his review of the literature on separation anxiety. He reviewed five studies in which three found no sex differences and two found that males evidence greater separation anxiety than females. In the most comprehensive review to date, Maccoby and Jacklin (1974) reached a similar conclusion. In several studies of children aged 10 months to 3 years, males exhibited greater resistance to separation, as indicated by greater distress at separation or greater likelihood of quickly following the departed figure. More recently Waters (1978) reported that the normative data for the Ainsworth "strange situation" indicated that crying and its correlates were greater in males than in females in the second separation and reunion sequence at the ages of 12 months and 18 months.

A group of studies that are cognate to the studies on separation and deprivation are the studies of adoptive children. This kinship resides in the fact that adoptive children experience a break in the continuity of care prior to placement and frequently come from deprived backgrounds (Hersov, 1977a). In an extensive review of this literature, Hersov concluded that these and other stresses result in adoptive children's being at greater risk for the development of a psychiatric disorder. Furthermore, in those few studies that have employed a control group, males have been found to be more at risk than females.

Further indication of greater male vulnerability to environmental stress is found in the cognitive realm. Bayley (1970) conducted a review on the association of parental variables such as warm, hostility, restrictiveness, and so on with IQ. She concluded that there were clear indications that the mental abilities of males are more strongly related than those of females, both positively and negatively, to the emotional aspects of their environment. She also noted that there were indications that early experiences are more likely to have long-lasting effects on boys than on girls. Similarly, in two separate reviews and a recent study, maternal employment has been found to have a negative effect on male academic achievement (Etaugh, 1974; Gold & Andres, 1978; Hoffman, 1974). And finally, in a literature review on the effects of father absence on children's cognitive development, Shinn (1978), although concluding that the proportion of studies that found negative effects was the same for males and females, noted that three studies found stronger negative effects for males.

Recented, however, Kamin (1978) reexamined Bayley's data and concluded that no significant sex difference in susceptibility of IQ to environmental influence has been demonstrated. Furthermore, he contended that the differences in male and female samples with regard to maternal education and children's IQ variance confound whatever sex difference may be detected. Thus, for the present, it appears that the significance of Bayley's findings has been attenuated.

Perhaps the most interesting aspect of the above finding is that it dovetails so nicely with the undisputed fact of greater male vulnerability to biological stressors. Hence one is tempted to posit a correspondingly greater constitutional vulnerability to psychological stressors. The data, though far less decisive than they are for biological stress, do offer some intriguing support for this hypothesis. For example, it may be that the same psychological stress is more severe for males than for females simply because it

interacts with a less mature organism. It may also be that just as the female organism is more stable for physical and mental growth (Mussen et al., 1974), so it is more stable in maintaining psychological stability.

However, there exist equally plausible explanations for the seemingly greater male vulnerability to psychological stress. For example, in the case of familial stress, what is apparently the same situation may in practice be different depending on the sex of the child. Thus it may be that the male child for some reason has more contact with the disturbed parent, has more responsibility for the home when the parent is ill and so on. Though this hypothesis was tested and found wanting by Rutter (1970), it has not yet been adequately explored. It may also be that boys and girls respond to different family stresses. For example, Hoffman (1974) noted that in most child developmental studies girls show ill effects from too much supervision or control, whereas boys typically suffer from too little. Further, if one examines the studies indicating greater male vulnerability to psychological stress, it seems that lack of control rather than too much control more correctly typifies the family situation. Hence, it may be that in future studies that examine a greater range of stress variables, females may prove to be more vulnerable than males to some of them.

Parenthetically, it should be noted that the reasons for differential effect of under- and overcontrol are not clear. Though it would be convenient to evoke the hypothesis of differential socialization of the sexes, Maccoby and Jacklin (1974) concluded that the data reveal a remarkable uniformity in the socialization of the sexes. This has been challenged by Block (1978) and Hoffman (1977), largely on the basis of more recent research, none of which addressed itself directly to the point at hand. One can only note that this finding has an interesting parallel in the clear sex difference in adult psychopathology, in which there exists greater male prevalence of disorders of undercontrol, that is, personality disorders, and a greater female prevalence of disorders of overcontrol, that is, neurotic disorders (Dohrenwend & Dohrenwend, 1974).

Learning Difficulty

There is a male preponderance in all disorders that involve a specific delay in development (i.e., speech or language delay, nocturnal enuresis, and clumsy child syndrome; Rutter, 1977a). This preponderance continues into the school years; Kessler (1966) noted that academic difficulty is the reason for referral of at least three fourths of the children between the ages of 7 and 14 and that it is indisputably a male problem. Whether one looks at mental retardation (Lehrke, 1972, 1978), reading difficulty (Rutter & Yule, 1977), hyperactivity (Cantwell, 1977), or simply lower grades (Maccoby & Jacklin, 1974), males predominate. Given the seriousness of this problem in its portent for future adult maladjustment (Kohlberg et al., 1972), an examination of its possible causes becomes of major significance.

The causes cited by Gove and Herb (1974) were the slower intellectual and physical development of the male and the incongruity of establishing a male identity in the feminine world of the school. Again it seems that the endowments and experiences of the sexes are more divergent than the causes noted by Gove and Herb.

The feminine culture theory of male learning difficulty, although confidently put forth by Gove and Herb as well as others, requires closer scrutiny as to exactly how it operates. For example, sex of teacher per se does not seem to be the prime factor, since a review of eight studies found no notable favorable effects of male teachers on male students (Good & Brophy, 1977). That male teachers have little, if any, differential influence on male achievement is explained by the fact that male teachers interact with boys and girls in the same general way that female teachers do (Good & Brophy, 1977).

More relevant feminizing factors than the sex of the teacher seem to be either what is studied or the cultural stereotype toward learning in general. For example, Good and Brophy reported two studies in which the sex

difference in reading was eliminated when males read material of high interest to them. Also, Johnson (1976) surmised that one reason why the sex difference in reading is not evident in cultures such as those of Nigeria, Germany, and England is that reading in these cultures is deemed a masculine activity. Furthermore, to the extent that these and other factors are operative, they seem to affect the grades teachers award to males more than they affect actual achievement. Thus Garai and Scheinfeld (1968), in their review of sex differences in intellectual performance, indicated that females in elementary and high school are generally awarded higher grades than males despite the fact that males achieve as well as or, in some cases, better than females.

Though these environmental factors undoubtedly account for part of the observed sex difference in some learning difficulties, there are some equally compelling biological factors stemming from the dimorphism of the sexes that play a significant role. The first and most obvious biological difference, which Gove and Herb (1974) mentioned, is that the male child at school age lags approximately 1 year behind the female in physical maturation (Garai & Scheinfeld, 1968). Tanner (1970) hypothesized that the male maturational lag is probably due, indirectly, to the action of the genes on the Y chromosome. Thus, children with the abnormal chromosome constitution XXY (Klinefelter's syndrome) have a skeletal maturity indistinguishable from that of the normal male, and children with the chromosome constitution XO (Turner's syndrome) have a skeletal maturity approximating that of the normal female XX constitution.

This greater female maturation appears to be paralleled by a greater intellectual maturation. Bayley (1956) indicated that physical growth, as measured by height and skeletal maturity, is positively correlated with IQ scores. Note, however, that in individual cases physical growth, as measured by percentage of mature height achieved, is not correlated with IQ measured in terms of percentage of 21-year-old intelligence scores achieved. Parenthetically it should be added here that Maccoby and Jacklin (1974) cited Bayley to caution against the acceptance of a correlation between maturation and intellectual growth. However, as Bayley clearly indicated, this lack of correlation pertains only to individuals and not to groups differing in maturation. Thus, Sherman (1978) aptly pointed out that though physical growth spurt might not correlate with mental growth spurt within individuals, groups could differ, with both the physical and the mental growth spurts coming earlier in females. Tanner, (1970, 1978), in his review of the relationship between physical maturation and mental ability, stated that the more physically mature scored higher on mental tests in North American and European populations at all ages tested, going back as far as 6½ years. He thus concluded that in age-linked examinations more physically mature children have a significantly better chance than less mature children.

At present there seems to be general agreement with Sherman's cautious statement that there is no easy resolution to the question of whether the sexes differ in cognitive maturation. Though, as she added, it would be unwise to ignore this possibility. This caveat seems to be well-founded and should be most heeded when one considers young children. For example, Maccoby and Jacklin (1974) noted that when sex differences are found in general intellectual abilities between the ages 2–7 years, these differences usually favor girls. Furthermore, Block (1976) pointed out that their caution in interpreting this finding as being due largely to the disadvantaged origin of the samples is unjustified.

Hence, it comes as no great revelation to find Anthony (1970) stating that in the first grade boys are referred for help 11 times as often as girls for social and emotional immaturity, a syndrome characterized by a high rate of absenteeism, fatigue, inability to attend and concentrate, shyness, poor motivation for work, inability to follow directions, slow learning, infantile speech patterns, and problems in the visual–motor and visual perception areas.

The second obvious biological difference, which has been discussed before, is the

greater male vulnerability to a host of pre-, peri-, and postnatal stresses resulting in brain dysfunction (Birns, 1976; Garai & Scheinfeld, 1968; Maccoby & Jacklin, 1974). Furthermore, the role of brain dysfunction in learning disturbance seems well established (Heincke, 1972; Rourke, 1975), though the precise nature of the dysfunction is vigorously disputed (Vellutino, 1977). Hence, one of the sequelae of the greater male vulnerability to biological stress is the greater prevalence of males who are predisposed to learning disorders.

The third more controversial biological difference is that of greater male variability. Shields (1975) in tracing the evolution of the concept noted that one of its first serious discussants was Darwin, who used it to explain how in many species males developed greatly modified sexual characteristics, whereas females did not. This principle was next brought to the attention of psychologists by Ellis (cited in Shields, 1975), who extended it to mental abilities as well as physical traits. After noting that there were more men than women in homes for the mentally deficient, which indicated a higher incidence of retardation among males, and that there were more men than women on the rolls of the eminent, which indicated a higher prevalence of genius among males, he concluded that greater male variability probably held for all qualities of character and ability. A current application of this principle to mental ability comes from Lehrke's (1972, 1978) X-linkage theory of intellectual traits.

Lehrke (1972, 1978), along with his predecessor Ellis, maintained that males are more frequently represented at the extremes of the range of general intelligence. He further proposed that this greater male variability is best explained by assuming that there are major genes for intelligence on the X chromosome. Concerning the first hypothesis of greater male variability at the extremes of intelligence, Lehrke's own review of the literature, along with an independent review of 27 community epidemiological surveys by Abramowicz and Richardson (1975), indicated a greater prevalence of male retardates in both institutions and communities, spread out in time over 80 years and in space from Australia to Scandinavia. His response to the criticism of sex bias in diagnosis as an explanation was as follows: First, he contended that if there is a bias, it is in the opposite direction, since retarded females are institutionalized more frequently as a means of controlling their fertility. Second, he maintained that it is hardly plausible that all the studies showed the same type of sex bias in determining which individuals were counted as retarded; and even if some such bias were present, he contended that it is hardly conceivable that it would result in such extreme levels as to account for the difference seen in most studies (e.g., as much as a 76% male excess in some studies).

However, he noted that those studies showing the greater numbers of retarded males may mean merely that there is a greater amount of pathology affecting males' intellectual functioning, unless it can also be shown that there are more males at the high end of the distribution of IQs and that the greater prevalence of males at the low end follows a sex-linked pattern of inheritance. Since the contention of a greater male prevalence is both seriously questioned (Maccoby & Jacklin, 1974; Sherman, 1978) and tangential to the present discussion, it is the evidence for the sex-linked pattern that is examined.

The most persuasive support for this linkage comes from the findings that retarded women married to normal men are twice as likely to have retarded offspring as retarded men married to normal women and that there is a marked excess of families with male-only retardates. Anastasi (1972) criticized these supporting data by suggesting that since women have more of a role in child rearing, retardation should have a more debilitating effect on their offspring than should the retardation of their husbands. The force of this criticism became blunted, however, when Lehrke noted that this social deprivation hypothesis applies mainly to mild retardation and can hardly explain a fivefold increase in severely retarded children born to retarded mothers as compared with retarded fathers. Furthermore, this criticism, along with the criticism that the major factor may be ma-

ternal prenatal rather than maternal gene influences on the X chromosome, fails to account for the strong tendency of male retardation to run in families and of which the most likely explanation is an X linkage. And finally, in a sample of 5,049 pairs of individuals, Freire-Maia, Freire-Maia, and Morton (1974) examined three alternate explanations of what they referred to as an established sex difference in retardation. They concluded that a sex-modified threshold for mental retardation that includes sex-linked genes is more consistent with the evidence than hypotheses that stress either prenatal environmental or postnatal maternal socialization.

Psychosexual Disorders

Davison and Neale (1978) indicated that in Diagnostic and Standard Manual III *sexual deviations* will probably become *psychosexual disorders* and be divided into three categories in which only the gender identity or role disorder category (e.g., transvestism and transsexualism) will be pertinent to a discussion of childhood disorders. The two other categories of paraphilias (e.g., fetishism and rape) and psychosexual dysfunctions (e.g., impotence and dyspareunia) are the province of adult psychopathology.

In an adult population males exceed females in disorders of gender identity (Green, 1974; Pauly, 1974); and though studies of this disorder in childhood are just beginning, what research does exist points to a similar greater male prevalence among children (Green, 1974). In childhood this disorder takes the form of a cross-gender identification on such dimensions as self-concept, playthings, clothing preference, playmate preference, and so on (Green, 1974). Although research on the significance of this behavior for females is virtually nonexistent, three follow-up studies on males have documented the continuity of this behavior into adulthood. Out of a total sample of 27, 15 became adult transsexuals, transvestites, or homosexuals (Green, 1977).

Two theories have been offered to explain this greater male prevalence. A psychoanalytically oriented theory of identification proposes that gender differentiation is more difficult for the male. For although both sexes initially identify with the mother, it is only the male who has to switch this identification (Green, 1974). Consequently, this additional hurdle entails the concomitant greater probability that male gender differentiation will not be successful.

A biologically based theory proposes that there is a prenatal hormonal hurdle that only the male has to surmount (Green, 1974): that is, as Money (1974) has indicated, it appears that nature's rule is that to masculinize something must be added. Specifically, at about the 6th week of prenatal development, masculinizing substances must be released by the fetal testis for morphologic differentiation of males, whereas no sex hormone whatsoever is necessary for the differentiation of morphologic females. Thus, if both embryonic gonads are removed prior to the critical period when the sexual anatomy is formed, then the embryo will proceed to differentiate as a morphologic female, regardless of genetic sex. Of these two masculinizing substances, one, known only by its function, is the müllerian inhibiting substance. Without it the male is born with the müllerian ducts differentiated into a uterus and fallopian tubes, as in the female. The other substance is androgen, the male sex hormone, essential for differentiation of the internal masculine reproductive tract and for the differentiation of the external sexual anlagen into male structures instead of their female homologues—that is, the clitoris, clitoral hood plus labia minora, and labia majora, instead of, respectively, the penis, penile skin covering, and scrotum. This release of fetal hormones results not only in the aforementioned genital dimorphism but also in a brain dimorphism, the significance of which is discussed later. Parenthetically, it should be mentioned that even in the last stages of differentiation in puberty, the changes of males are more striking and extensive than those of females (Tanner, 1978).

Thus, in a dual system in which the female path automatically evolves and the alternate male path requires specific influences at spe-

cific intervals, more errors probably occur along the latter path. This greater probability of error is theorized to contribute to the greater prevalence of male gender identity disorders.

Antisocial Behavior

In their review Kohlberg et al. (1972) stated that antisocial behavior, particularly when some estimate of severity is taken into account, is the single most powerful predictor of adult maladjustment. Hence the decisive male preponderance in aggressive behavior, which is probably the most unequivocal sex difference in the literature (Feshbach, 1970; Hoffman, 1977; Maccoby & Jacklin, 1974; Terman & Tyler, 1954), takes on a special significance. Furthermore, this difference, as Maccoby and Jacklin concluded, is real and cannot be explained away by simply positing different modes of expressing aggression (Feshbach, 1970). However this should not be construed to mean that women are always less aggressive than men (Frodi, Macaulay, & Thome, 1977). This preponderance in aggressive behavior manifests itself in males' exceeding females in both delinquent and nondelinquent disturbances of conduct (Wolff, 1977) and in externalizing symptoms in general (Anthony, 1970). This ratio persists into adulthood, where males predominate in personality disorders (Dohrenwend & Dohrenwend, 1969, 1974).

The most common explanation given for this disparity is differential socialization. Mead (1949) provided some of the most dramatic data for this explanation with her findings of sex role reversal among the Tchambuli, in which the females were the aggressive, dominating personalities. Parenthetically, it should be noted that the reversal was not complete, since it is the men who fight when the Tchambuli go to war (Brown, 1965). Since then authors, following the lead of Sears, Maccoby, and Levin's (1957) finding that parents make the greatest distinction between child rearing of boys and girls in the area of aggression, have attempted to investigate the nuances of this socialization process. Reviews on sex role acquisition by Mischel (1966), Mussen (1969), Feshbach (1970), Bardwick (1971), and Maccoby and Jacklin (1974), just to name a few, typically mention learning mechanisms such as operant conditioning, modeling, and frustration aggression as mediators of the socialization process. On the basis of these and numerous other reviews, further slaying of the Freudian myth of innate female passivity would be tedious overkill; and one can safely assume that differential socialization is a major factor in explaining the sex differences in antisocial behavior.

As with many myths, however, there resides an element of truth; and it is likely that differential socialization is not the complete explanation. Most authors have recognized this. For example, Feshbach (1970), in his exhaustive review of childhood aggression, wondered "whether from a biosocial view, it is also reasonable to ask whether it is easier to facilitate aggressive behaviors in boys than in girls and what the implications of this training might be for other behaviors of the child" (p. 189). In view of current research, it might be added that it would be unreasonable to assume otherwise. In the most extensive review to date, Maccoby and Jacklin (1974) concluded that the higher level of male aggression probably cannot be completely explained by a learned fear of aggression among girls, by any tendency for girls to reinforce the aggression of boys, or by the tendency of adults to reinforce aggression more in males. This conclusion regarding the socialization of aggression is part and parcel of Maccoby and Jacklin's overall conclusion that parental socialization data show little differences between the sexes. And although this surprising and controversial interpretation has been seriously questioned (Block, 1978), the conclusion that differential socialization is not an exhaustive explanation of the sex difference in aggression seems secure, as the following discussion soon indicates. Maccoby and Jacklin contended that the male's greater aggression has a biological component. This biological disposition to greater aggression stems not only from the greater mesomorphy of the male,

which is manifested by age 5 (Willerman, 1979), but also from the effects of the male hormone on the organism.

The correlation between mesomorphy and variables such as aggression and delinquency is clearly established whether one interprets this correlation as a manifestation of the same underlying biological structure or as the influence that different constitutions have on the successful reinforcement of instrumental responses (Feshbach, 1970; Hall & Lindzey, 1973; Shah & Roth, 1974). This predisposition to aggression because of greater psysical strength is further reflected in the fact that there is no society on record in which the female does the actual fighting in warfare (Brown, 1965).

The second factor predisposing the male to greater aggressivity is the male hormone, androgen, whose importance has been dramatically demonstrated in animal research. Reviews by Money and Ehrhardt (1972), Hart (1974), Quadagno, Briscoe, and Quadagno (1977), Reinisch and Karow (1977), and Vandenberg (1978) indicate that whereas the male of the vertebrate species is the more aggressive in both laboratory and natural situations, the administration of the male hormone either pre- or postnatally results in the female's being equally aggressive.

Something analogous to this also occurs in humans. Block (1976) disagreed with Maccoby and Jacklin's (1974) verdict of no sex differences in activity level. She pointed out that not only have they erred in their interpretation of some of the studies but they have also omitted nine relevant studies, all of which reported a higher male activity level. In addition, a recent epidemiological investigation of 705 3-year-olds randomly sampled from the community reported that although the overall prevalence of behavior problems did not differ for the sexes, males did exceed females in being described as too active (Richman et al., 1975). Furthermore, Willerman (1979) indicated that mortality figures among children aged 1–4 years show that boys are much more likely than girls to die from accidents and that many of these differences are clearly present around age 1. Although acknowledging the possibility of differential child-rearing practices, he concluded, and is supported in this conclusion by Maccoby (1976), that there are only weak indications that parents may be less watchful of boys than girls in the early years. Hence, even though Fagot (1978) has recently reported that in children aged 20–24 months, males are more likely to be left alone by their parents, it seems that Willerman is still correct in concluding that boys are more likely to have a higher activity level than are girls.

Hence, a higher male activity level (which serves as a precursor to a higher aggression level [Patterson, Littman, & Bricker, 1967]) seems to be an established sex difference. Furthermore, Quadagno et al.'s recent review of the research on fetally androgenized human females supports Maccoby and Jacklin's contention that the consistency of findings with animal experimental work of a higher activity level among females is especially compelling in establishing a biological base for this difference.

This parallel with the animal research is all the more convincing in light of the excessive caution that Maccoby and Jacklin attached to attributing too much significance to this analogue. They, along with Quadagno et al., suggested that the cortisone therapy that the females received may account for their higher activity level. However, this caution is rendered less persuasive when it is noted that the females whose syndrome was progestin induced did not receive such treatment and yet manifested the higher activity levels. Furthermore, similar investigations of boys with either the adrenogenital or progestin-induced syndromes reveals comparable increases in activity level (Brecher, 1971; Erhardt & Baker, 1975).

They also warned that since the results were based on parental report, the report may have been inaccurate or the parental perception of the females as more tomboyish may itself have induced the reported difference; but this also seems overly cautious. Reinisch and Karow (1977), in their review of the effects of prenatal exposure to synthetic estrogens and progestins on human development, concluded that it seems unlikely that

the fact a mother knew her child was treated had a significant effect on her treatment of that offspring. They indicated that different hormone treatments have been shown to have different effects, which is difficult to explain in terms of maternal caretaking.

There is additional evidence that maternal knowledge is not a significant factor in the common finding that mothers have no recall of their having taken the hormones. Even in the case in which the effects are a change in genital morphology, as in the oft quoted Erhardt and Money (1967) study, parental attitude toward the child and her behavior is difficut to predict. Erhardt (1973) noted that parents did not have a consistent attitude toward their masculinized daughter, when compared with control parents, that could explain the tomboy syndrome.

Neurosis

In contrast with children with antisocial symptoms, children with neurotic problems are almost as likely to be mentally healthy adults as a random sample of the population (Hersov, 1977b; Kohlberg et al., 1972; Rutter, 1972b). Also in marked contrast with the female preponderance in adult neurosis (Dohrenwend & Dohrenwend, 1969, 1974), males either equal or exceed females in childhood neurosis (Gilbert, 1957; Gove & Herb, 1974; Hersov, 1977b). It is not until adolescence that the adult ratio begins to manifest itself (Gove & Herb, 1974; Hersov, 1977b; Rutter, 1974; Terman & Tyler, 1954). Many of the reasons for this greater male prevalence seem to be similar to those already adduced for the greater male prevalence in childhood adjustment reactions and hence are not in need of further elaboration. Furthermore, with the additional explanation of adverse reactions secondary to the greater male prevalence of learning problems (Corbett, 1977; Rutter & Yule, 1977), it seems that the major interpretations have been explored. Hence, what remains is to examine the reasons for the difference between child and adult sex ratios.

Gove and Herb (1974) concisely summarized the explanations that rely on sex role theory. These explanations are evaluated in the light of more current research and complemented by explanations that have a more biological focus.

Gove and Herb sounded their familiar theme of differential stress and hypothesized that adolescence brings an increase of stress for females and a decrease for males. The feminine sex role is thought to become more stressful first because there is a sudden narrowing of the sex role, in that the female is restricted from engaging in activities that are then deemed too masculine. Such a precipitous constriction of sex role is thought to induce conflict and anxiety since these activities have most likely been integrated into her personality. The prime example given of this type of stress is that females who were once rewarded for academic success find, in adolescence, that they should not surpass men. Consequently, they come to fear success. In support of this hypothesis, Hoffman (1977) indicated that females, though no less achievement oriented than males, appear more attuned to the negative consequences of academic and occupational success. This concern reflects the realities of the situation, since for women—particularly during the adolescent and college years when heterosexual relations are salient—the rewards of high academic and occupational success are uncertain and their costs—often in the form of affiliative loss—real. However, if fear of success is assessed on a fantasy basis, there appears to be little overall sex difference (Tresemer, 1974; Zuckerman & Wheeler, 1975).

Second, females are hypothesized to perceive their roles as depending on the actions of others, and because of this they experience more uncertainty about the future. This hypothesis receives strong support from Douvan and Adelson's (1966) massive study of 3,500 adolescents in the 1950s. Both then and now, for a majority of females identification with a future adult role involves primarily that of wife and mother (Conger, 1977). As a result, Douvan and Adelson concluded that females' adolescent adaptation is more difficult because they face a more ambiguous task in adapting their present life and self-concept to the future. The ambi-

guity stems from the fact that marriage, unlike occupation, is less a matter of simple individual choice for females than for males and lies not in the immediate future but beyond in some relatively unspecified time. Marriage lends itself neither to rational planning nor to specific preparation, since it involves the decision and initiative of another person.

Furthermore, it seems that with the traditional feminine role in a state of transition, the assumption of a future feminine identity is becoming even more ambiguous. The alternate satisfactions of career achievement are often seen as incompatible with those of marriage and motherhood (Hoffman, 1977). As Conger suggested, this may mean that the female is exposed to conflicting social rewards and punishments no matter which role she assumes.

Third, because of their characteristic dependency, females are hypothesized to find the transition to an independent status, which starts in adolescence, more difficult. Though Maccoby and Jacklin (1974) have challenged this conventional notion of greater female dependency, other reviewers have reaffirmed its validity (Block, 1976; Hoffman, 1977). What appears to be disconfirmed, however, is the hypothesis that females find the transition to an independent status more stressful. Conger's review indicates that females appear to experience fewer and less stressful conflicts over the development of independence than do males, particularly in early adolescence. He indicated that males are more likely to be actively engaged in establishing independence from parental control, to be concerned with issues of self-esteem and achievement of responsibility for their own actions, and to be more preoccupied with issues of self-control (e.g., control of temper and impulsiveness).

It may be possible to reconcile these seemingly contradictory positions in two ways. First, it may be that Conger's (1977) findings are applicable to late adolescence. This possibility however, while tempting, remains speculative. A second possibility with more research support is that the two positions look at different manifestations of the same phenomenon. Rutter, Graham, Chadwick, and Yule (1976) examined the concept of adolescent turmoil in the context of findings from a total-population epidemiological study of Isle of Wight 14–15-year-olds. Though their findings do not directly relate to the independence issue, an extrapolation to this issue seems appropriate. The study concluded, along with American studies as summarized by Conger, that parent–child alienation is not a common feature unless adolescents have already showed psychiatric problems. In addition, the study supported Conger in finding that when alienation existed, it was more common among males. However, the study also concluded that inner turmoil as represented by feelings of misery and self-depreciation was quite frequent and more common among females. Thus it may be this manifestation of stress that Gove and Herb have alluded to.

The fourth and final hypothesis offered by Gove and Herb is that as adolescent females prepare for and start moving into adult roles they become aware that males are favored in our society and start experiencing the stress associated with their adult sex role. Though this hypothesis is offered more in the form of a confident assertion than as a result of well-based knowledge, a study by Peskin (1972) confirms its basic thrust. Using a sample of 31 male and 33 female subjects from the longitudinal Berkeley guidance study (cited in Peskin, 1972), he reported that for both sexes a change from a relatively tension-free or conflict-free preadolescence to a stressful adolescence is predictive of positive adult mental health. He further found that a greater number of these changes, for example, from high to low self-confidence, are predictive of positive adult mental health for females than for males. Peskin concluded that adolescence clearly appears more disquieting for the female, possibly because of her unique task of acquiring what psychoanalytically oriented theorists term a *self-assured passivity*.

In addition to these hypotheses, there are others that focus on two of the basic biological differences between the sexes, the internal and external sexual structures. These

dimorphisms tend to be construed by Freudian theorists in terms of biological imperatives with pervasive influences on psychosexual functioning. Other theorists propose that the reverse is likely to be true and that the sexual area may be precisely that realm in which the superordinate position of the sociocultural over the biological level is most convincing (Gagnon & Simon, 1973). It is the implications of this latter position that are sketched in detailing additional biologically based reasons for a more stressful female adolescence.

With the onset of female adolescence, there occurs the sexually dimorphic trait of a cyclic release of gonadotrophin, which results in, among other things, menstruation. Furthermore, it appears that for many women there are also attendant feelings of tension, irritability, anxiety, and depression, composing the syndrome of premenstrual tension (Bardwick, 1971; Conger, 1977). This syndrome has been found to be correlated with violence, death by accident and suicide, and admission to a hospital because of acute psychiatric disorder (Graham & Rutter, 1977; Moyer, 1973). Although it is not clear to what extent this syndrome can be attributed to actual physiological causes or to the script of a culturally induced expectation (Parlee, 1973; Ruble, 1977), it is commonly accepted that adverse emotional consequences are a corollary for at least some women (Graham & Rutter, 1977). And to the extent that this is so, it can be considered contributory to a more stressful female adolescence.

A second hypothesis focuses on the reactions of adolescents to the physical changes engendered by puberty. Conger (1977) reported that such changes cause more concern for females, with a substantial minority reporting feeling shame and anxiety over a change such as menstruation. This greater female concern is due in part to the fact that only the female has to face the possibility of an unwanted pregnancy. This possibility is the number one worry of parents of females (Hoffman, 1977) and is undoubtedly communicated to their daughters. Consequently, whereas the male has only to fear sexual failure, the female has to fear failure and success simultaneously (Gagnon & Simon, 1973). Hence it is not surprising that the most common reaction of females to the first experience of intercourse is fear, whereas for males it is excitement (Sorenson, 1973).

In addition, the physical changes with their dramatic effects on physical attractiveness are a further cause of the greater concern of females. Berscheid and Walster (1975) have amply documented the overriding importance of physical attractiveness in the adolescent dating situation, and this probably explains why physical attractiveness outweighs all other concerns in early adolescence (Eme & Goodale, in press). Furthermore, since female popularity is more closely related to physical attractiveness (Berscheid & Walster, 1975), it comes as no great revelation that physical attractiveness causes them more concern than it does males (Eme & Goodale, in press; Musa & Roach, 1973).

In sum, these hypotheses suggest that, beginning in adolescence, females tend to experience more stress than males. Moreover, they seem more disposed to respond to this stress according to the sex-stereotypic pattern of internalization rather than externalization (Achenbach, 1966; Anthony, 1970; Garai, 1970; Locksley & Douvan, Note 1), with the resultant greater prevalence of neurotic symptomatology.

These preceding hypotheses that suggest a more stressful female adolescence are complemented by others that suggest a lessening of stress in some areas for males. Gove and Herb (1974) indicated that males start performing better academically. This is attributed to a maturational "catch-up," the increasing importance of "masculine" subjects such as mathematics and science and the increased relevance of schooling to their long-range vocational goals. These hypotheses, like the previous ones, are evaluated in the light of more current research and are complemented by explanations that have a more biological focus.

The hypothesis of maturational catch-up receives support from the most extensive review to date on sex-related cognitive differences, that by Sherman (1978). She endorsed the generalization that though females are

generally verbally precocious compared with males, males eventually catch up, although females retain a slight edge in verbal functioning. She attributed this catch-up both to maturation and to the exposure of males to heavy educational intervention in verbal training.

In the area of mathematical skills, Sherman, along with Maccoby and Jacklin (1974), concluded that beginning in adolescence males tend to outperform females. She differed from Maccoby and Jacklin by placing the onset later in adolescence, by assessing the difference as nonexistent or minimal once previous mathematical background was controlled for, and by rejecting a biologically based male superiority in spatial ability as contributory to whatever male superiority did exist. However, reviews by Waber (1977) and Goleman (1978) have both affirmed the strong possibility of a biological basis for greater male spatial ability.

Despite the preceding disagreements, it does seem clear that whether one views male adolescent mathematical ability as greatly or minimally superior to that of females, it undoubtedly contributes to improving male academic performance. Also, there is sufficient support in the data for a biological basis for this ability, Sherman notwithstanding, and at the very least, a prudent person would not totally ignore this possibility, as did Gove and Herb (1974).

The increased relevance of schooling to the long-range goals of males needs to be closely scrutinized in light of the significant changes in the vocational orientation of females (Conger, 1977). Although this greater relevance was undoubtedly true some 20 years ago when the studies cited by Gove and Herb, such as that of Douvan and Adelson (1966), were actually conducted, there is a question as to its continuing validity. For example, in the most thorough study to date, Buxton (1973) surveyed 6,500 students in grades 7–12 in four separate school systems and reported that both sexes were highly concerned with the relation of schooling to the future. Moreover, females were found to deny disliking school more strongly than males and to like teachers more and to express more feelings of guilt and anxiety about school. Hence it may be that schooling does take on increased relevance for males, but this does not mean that it becomes less relevant for females. What may be true, however, is that despite the fact females are at least equally concerned about school as males, this concern does not have the same consequences for them. Thus, Locksley and Douvan (Note 1) reported that objective academic achievement was associated with less stress for adolescent males but not for females. Indeed, they found that females with high grade point averages were more depressed and reported more psychosomatic symptoms than males with high grade point averages, and they were not significantly less aggressive than females with low grade point averages. Locksley and Douvan suggested that the reason why actual academic achievement did not reduce stress in females is that grades constitute a basis for social comparison with peers, precipitating conflicts over standards of femininity and sexual desirability, and that the anticipated ramifications of academic achievement for future work and family plans are conflictful for females. To the extent that this is so, it reflects Garai's (1970) conclusion that for males happiness appears to be highly correlated with some strong vocational interest that manifests itself frequently during puberty or even earlier, whereas females are more interested in interpersonal relations. Consequently, though commitment to vocational goals, as signified by successful academic achievement, may serve to mitigate the stress of adolescence for males, for females such success may paradoxically precipitate even more stress.

Although the preceding discussion offers support for a decrement in stress in adolescence for males relative to females, there is other evidence that suggests that this decrement may be somewhat illusory. Because of the cultural proscription against male emotionality (Hoffman, 1977), males may learn to mask the neurotic expression of their problems in much the same way that Kagan and Moss (1962) have suggested they mask the expression of dependency. Toolan (1962) and Glaser (1967) have both indicated that

many of the acting-out symptoms of adolescents may camouflage underlying depression. This may explain why Achenbach and Edelbrock (1978), in one of the most sophisticated factor analytic studies to date of syndromes derived from parental report of clinic-attending children (aged 6–16 years), indicated that all samples except the 12- to 16-year-old male sample, yielded a factor labeled depression. It is not surprising, then, that Graham and Rutter (1977) indicated that whereas boys with disturbances of conduct are likely to be referred to a psychiatrist, when they get older these same personality disturbances or criminal behaviors are often likely to be dealt with in other ways. Thus the altered sex ratio in neurotic disorders that begins in adolescence and with it the implication of a less stressful male adolescence may be partly an artifact of a different manifestation of stress and the consequent referral process.

Furthermore, the increment in stress in adolescence for females, as indexed by greater "inner turmoil" (Rutter et al., 1976), may also be partly illusory. Garai (1970) suggested that the presence of anxiety or fear may not be a reliable indicator of mental disturbance for females. He noted that since they seem to live at a generally higher level of anxiety than males do, this may be a sign of alertness to danger and stress and may enable them to cope better with emergencies and endure anxiety and fear for longer periods of time than men, who tend to repress anxiety. Hence to a certain extent greater female inner turmoil may be adaptive rather than maladaptive.

In summary, whereas in childhood it seems that cultural pressures are more dissonant for male predispositions, in adolescence they are more dissonant for females. This dissonance can be expected to be coped with in sex-stereotyped patterns of internalization and externalization, with the correlative change in the sex ratio for neurotic disorders.

Psychosis

For a long time most writers tended to group all the psychoses of childhood together, usually under the term *schizophrenia of childhood*. However, in recent years the situation has changed radically, and there is a general (but not quite universal) recognition that the generic term *childhood schizophrenia* covers a number of quite different syndromes, which should be differentiated (Rutter, 1977c). The distinction that is most commonly made is between autism and childhood schizophrenia. This distinction has received impressive support (Rutter, 1978) and is embraced in this presentation for reasons that are detailed later. However, it should be noted that many, such as Bender (1971) and Miller (1974), reject this position; and in fact, the National Society for Autistic Children has adopted the official posture that to conceptualize autism as the earliest form of schizophrenia, which becomes manifested in later childhood or early adulthood, is equally valid. Though agreement on this distinction is lacking, there is enough specificity to at least distinguish nonpsychotic from psychotic children (Keith, Gunderson, Reifman, Buschbaum, & Mosher, 1976). In addition, there is agreement that whatever the syndromes of child psychosis may be, it is the most debilitating of all forms of psychopathology in the severity of its symptoms and the bleakness of its prognosis (Hintgen & Bryson, 1972; Rutter, 1977c).

With regard to autism, there seems to be universal agreement that males outnumber females in ratios ranging from 2:1 to 4:1 (Goldfarb, 1970; Hintgen & Bryson, 1972; Kanner, 1973; Rutter, 1978; Schopler, 1978). This ratio differs from that of adult schizophrenia, in which the sex ratio is the same (Dohrenwend & Dohrenwend, 1969, 1974; Rosenthal, 1970) and is one of the facts adduced to advance the contention that autism and adult schizophrenia are different nosological entities.

Although there is little agreement on the specific causes of autism, there is general agreement among those who have reviewed the research that when the etiology is eventually established, organic factors will be found to play a major role (Ornitz & Ritvo, 1976; Rutter, 1977c; Schopler, 1978). The one reviewer (Ward, 1970) who opted for an

etiology of primarily psychological factors seems to have been seriously deficient in his coverage of the literature (L'Abate, 1972; Rimland, 1972). As the most current reviews indicate, the majority of autistic children demonstrate severe deficits in intellectual, perceptual, and linguistic development (Friedman, 1974; Gunderson, Autry, & Mosher, 1974; Hintgen & Bryson, 1972; Rutter, 1977c; Rutter, 1978). Accordingly, it is currently thought that deficits in social behavior are less a cause than a reflection of dysfunction in other areas of development.

If one adopts this position, then it seems that the sex ratio can be explained largely in the same terms as were used to explain the similar sex ratio in learning difficulties. For example, one of the more tenable interpretations of childhood autism is that it is explicable primarily in terms of a language deficit, which in turn is rooted in defective cognitive functioning (Rutter, 1977c). This hypothesis is remarkably similar to Vellutino's (1977) conceptualization of dyslexia. In a comprehensive review, he rejected explanations based on visual perception, intersensory integration and temporal-order processing and concluded that the evidence most favors a verbal-deficit hypothesis.

One must add the caveat, however, that this analogy does limp quite noticeably, as Rutter (1968, 1974) pointed out. First, two of the most conspicuous contributors to learning difficulties, retardation and central nervous system dysfunction, are absent in fully one fourth to one third of autistic children. Second, the vast majority of children with learning difficulties do not manifest the symptoms common to autism. Furthermore, even though the learning difficulties concomitant with retardation characterize the majority of autistic children, it is clear that autism constitutes a syndrome that is different from that of retardation (Rutter, 1978; Schopler, 1978). Third, even where the analogy is most apt, as in the case of children with developmental, receptive-language disorders, autistic children differ in several important ways (Rutter, 1977c). Hence, until a more exotic kind of learning difficulty is discovered, the parallel, although currently considered the most plausible, remains imperfect.

As noted before, the greater male prevalence in autism stands in contrast with the parity of the sex ratio in adult psychosis and hence requires explanation. The most compelling explanation involves the medical model and posits different disease entities. The affective psychoses, which are virtually unheard of in childhood (Lefkowitz & Burton, 1978; Rutter, 1977c), are thought to have a strong biological base (Becker, 1977; Depue & Monroe, 1978; Gershon, Bunney, Leckman, Van Eerdewegh, & DeBauche, 1976) and this base may simply program a vulnerable organism for an adult rather than a childhood manifestation and for the resultant sex ratio.

Likewise, Rutter (1977c, 1978) made a strong case for autism's being a disease entity different from schizophrenia. He noted that autism differs from schizophrenia in terms of time of onset, social class, family history of schizophrenia, evidence of cerebral dysfunction, symptom patterns, level of intelligence, and course of the disorder. This distinction is further buttressed by the impressive evidence for genetic influences in schizophrenia (Cancro, 1976; DeFries & Plomin, 1978; Folstein & Rutter, 1977; Gottesman & Shields, 1976) in contrast with the minimal evidence for autism (Folstein & Rutter, 1977; Hanson & Gottesman, 1976). Thus, as with the affective psychoses, the sex differences in autism and schizophrenia can be most plausibly explained by positing different biological factors that result in different disease entities and hence in different sex ratios.

Although the preceding discussion has conceptualized autism and adult psychosis as distinct disease entities and has employed this distinction to explain the different sex ratios, mention was also made of the fact that a psychotic process similar to adult schizophrenia can manifest itself in childhood. This manifestation also involves a preponderance of males (Rutter, 1977c); but since it is conceptualized as being similar to adult schizophrenia, an explanation other

than one based on a biological difference is required.

The most plausible explanation stems from research on sex differences in the child and adolescent preschizophrenic personality. Both follow-back studies (Watt & Lubensky, 1976; Watt, Stolorow, Lubensky, & McClelland, 1970; Woerner, Pollack, Rogalski, Pollack & Klein, 1972) and preliminary findings of ongoing high-risk studies of schizophrenia (Mosher, Gunderson, & Buschbaum, 1972) indicate that whereas female preschizophrenics present a pattern of overinhibition, sensitivity, conformity, and introversion, males, in contrast, present patterns of unsocialized aggression. Hence it may be that because the unsocialized aggressive symptoms of males are more salient, their schizophrenia is more easily recognized, and therefore they are more likely to be diagnosed than female schizophrenics.

Summary and Conclusions

Recognizing the absence of a generally accepted taxonomy for use in diagnosing psychopathology in children, the present review chose to examine sex differences by focusing only on those categories relevant to the most important issue of continuity between child and adult psychiatric disorders. The review revealed a greater male prevalence in all of the following categories: adjustment reactions, antisocial disorders, gender identity disorders, learning disorders, neurotic disorders, and psychotic disorders. This finding stands in marked contrast with adult sex differences, which, beginning in adolescence, eventuate in a greater female prevalence in neurotic disorders and affective psychotic disorders, a greater male prevalence in personality and gender identity disorders, and no sex difference in schizophrenic disorders. The reasons for the childhood sex ratios and their lack of continuity into adulthood were examined in light of what is currently known about the differential endowment and experiences of the sexes.

This contrast between child and adult sex differences in psychopathology is perhaps the most salient finding to emerge from the review and suggests the two following conclusions. First, the validity of the contrast relies to a significant extent on the use of referrals rather than community surveys to arrive at the childhood sex differences. To the extent that this obviously limited source is validated in other ways, it seems apparent that the male child is more at risk for maladjustment than the female. Hence it would be wise for those engaged in primary prevention programs to take cognizance of this and address their efforts accordingly.

Second, though differential stress is clearly a major factor in explaining the contrast, it is equally clear that sex role per se is not the exclusive mediator of this stress. Biological factors play an important role and need to be given far more attention than they have been given in the recent past.

Reference Note

1. Locksley, A., & Douvan, E. *Problem behavior in adolescents.* Unpublished manuscript, University of Michigan (Ann Arbor), 1977.

References

Abramowicz, M., & Richardson, S. Epidemiology of severe mental retardation in children: Community studies. *American Journal of Mental Deficiency,* 1975, *80,* 18–39.

Achenbach, T. The classification of children's psychiatric symptoms: A factor analytic study. *Psychological Monographs,* 1966, *80*(7, Whole No. 615).

Achenbach, T., & Edelbrock, C. The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin,* 1978, *85,* 1275–1301.

Anastasi, A. Four hypotheses with a dearth of data: Response to Lehrke's "A theory of X-linkage of major intellectual traits." *American Journal of Mental Deficiency,* 1972, *76,* 620–622.

Anthony, J. Behavior disorders. In P. Mussen (Ed.), *Carmichael's manual of child psychology.* New York: Wiley, 1970.

Bardwick, J. M. *Psychology of women: A study of bio-cultural conflicts.* New York: Harper & Row, 1971.

Battle, E., & Lacey, B. A. Context for hyperactivity in children over time. *Child Development,* 1972, *43,* 757–772.

Bayley, N. Individual patterns of development. *Child Development,* 1956, *27,* 45–74.

Bayley, N. Development of mental abilities. In P. Mussen (Ed.), *Carmichael's manual of child psychology.* New York: Wiley, 1970.

Becker, J. *Affective disorders.* Morristown, N. J.: General Learning Press, 1977.

Bender, L. Alpha and omega of childhood schizophrenia. *Journal of Autism and Childhood Schizophrenia,* 1971, *1,* 115–118.

Berscheid, E., & Walster, E. Physical attractiveness. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 8). New York: Academic Press, 1975.

Birns, B. The emergence and socialization of sex differences in the earliest years. *Merrill-Palmer Quarterly,* 1976, *22,* 229–250.

Block, J. H. Issues, problems, and pitfalls in assessing sex differences: A critical review of *The psychology of sex differences. Merrill-Palmer Quarterly,* 1976, *22,* 283–308.

Block, J. Another look at differentiation in the socialization behaviors of mothers and fathers. In F. Denmark & J. Sherman (Eds.), *Psychology of women: Future directions of research.* New York: Psychological Dimensions, 1978.

Bowlby, J. *Attachment and loss: II. Separation.* New York: Basic Books, 1973.

Braine, M. D., Heimer, C., Wortis, H., & Freedman, A. M. Factors associated with impairment of the early development of prematures. *Monographs of the Society for Research in Child Development,* 1966, *31*(4, Serial No. 106).

Brecher, E. *The sex researchers.* New York: New American Library, 1971.

Brown, R. *Social psychology.* New York: Free Press, 1965.

Buxton, C. *Adolescents in the schools.* New Haven, Conn.: Yale University Press, 1973.

Cancro, R. Genetic and environmental variables in schizophrenia. In A. Kaplan (Ed.), *Human behavior genetics.* Springfield, Ill.: Charles C Thomas, 1976.

Cantwell, D. Hyperkinetic syndrome. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977.

Chesler, P. *Women and madness.* New York: Avon Books, 1972.

Chess, S., & Thomas, C. Differences in outcome with early intervention in children with behavior disorders. In M. Roff, L. Robins, & M. Pollack (Eds.), *Life history research in psychopathology* (Vol. 2). Minneapolis: University of Minnesota Press, 1972.

Conger, J. *Adolescence and youth.* New York: Harper & Row, 1977.

Conger, A. J., & Cole, J. D. Who's crazy in Manhattan: A reexamination of "Treatment of psychological disorders among urban children." *Journal of Consulting and Clinical Psychology,* 1975, *43,* 179–182.

Corbett, J. Mental retardation: Psychiatric aspects. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977.

Davison, G., & Neale, J. Abnormal psychology: An experimental clinical approach. New York: Wiley, 1978.

DeFries, J., & Plomin, R. Behavior genetics. In M. Rosenzweig & L. Porter (Eds.), *Annual review of psychology* (Vol. 29). Palo Alto, Calif.: Annual Reviews, 1978.

Depue, R. A., & Monroe, S. M. The unipolar-bipolar distinction in the depressive disorders. *Psychological Bulletin,* 1978, *85,* 1001–1029.

Dohrenwend, B., & Dohrenwend, B. *Social status and psychological disorder.* New York: Wiley, 1969.

Dohrenwend, B., & Dohrenwend, B. Social and cultural influences on psychopathology. In M. Rosenzweig & L. Porter (Eds.), *Annual review of psychology* (Vol. 25). Palo Alto, Calif.: Annual Reviews, 1974.

Dohrenwend, B., & Dohrenwend, B. Sex differences and psychiatric disorders. *American Journal of Sociology,* 1975, *80,* 1447–1454.

Douvan, E., & Adelson, J. *The adolescent experience.* New York: Wiley, 1966.

Eme, R., & Goodale, W. Seriousness of adolescent problems. *Adolescence,* in press.

Erhardt, A., & Baker, S. Hormonal orientations and their implications for an understanding of normal sexual differentiation. In P. Mussen, J. Conger, & J. Kagan (Eds.), *Basic and contemporary issues in developmental psychology.* New York: Harper & Row, 1975.

Erhardt, A. Maternalism in fetal hormonal and related syndromes. In J. Zubin & J. Money (Eds.), *Contemporary sexual behavior: Critical issues in the 1970s.* Baltimore, Md.: Johns Hopkins University Press, 1973.

Erhardt, A., & Money, J. Progestin-induced hermaphroditism: I.Q. and psychosexual identity in a study of ten girls. *Journal of Sex Research,* 1967, *3,* 83–100.

Etaugh, C. Effects of maternal employment on children. *Merrill-Palmer Quarterly,* 1974, *2,* 71–98.

Fagot, B. The influence of sex of child on parental reactions to toddler children. *Child Development,* 1978, *49,* 459–465.

Feshbach, S. Aggression. In P. Mussen (Ed.), *Carmichael's manual of child psychology* (Vol. 2). New York: Wiley, 1970.

Folstein, S., & Rutter, M. Genetic influences and infantile autism. *Nature,* 1977, *205,* 726–728.

Freire-Maia, A., Freire-Maia, C., & Morton, N. Sex effect on intelligence and mental retardation. *Behavior Genetics,* 1974, *4,* 269–272.

Friedman, E. Early infantile autism revisited. *Journal of Child Clinical Psychology,* 1974, *3,* 4–10.

Frodi, A., Macaulay, J., & Thome, P. Are women always less aggressive than men? A review of the experimental literature. *Psychological Bulletin,* 1977, *84,* 634–660.

Gagnon, J., & Simon, W. *Sexual conduct: The social sources of human sexuality.* Chicago: Aldine, 1973.

Garai, J. Sex differences in mental health. *Genetic Psychology Monographs,* 1970, *81,* 123–142.

Garai, J., & Scheinfeld, A. Sex differences in mental

and behavioral traits. *Genetic Psychology Monographs,* 1968, *77,* 169–299.

Gershon, E. S., Bunney, W., Leckman, J., Van Eerdewegh, M., & DeBauche, B. The inheritance of affective disorders: A review of data and of hypotheses. *Behavior Genetics,* 1976, *6,* 227–261.

Gilbert, G. A survey of referral problems in metropolitan child guidance centers. *Journal of Clinical Psychology,* 1957, *13,* 37–42.

Glaser, K. Masked depression in children and adolescents. *American Journal of Psychotherapy,* 1967, *21,* 565–574.

Gold, D., & Andres, D. Developmental comparisons between ten-year-old children with employed and nonemployed mothers. *Child Development,* 1978, *49,* 75–84.

Goldfarb, W. Childhood psychosis. In P. Mussen (Ed.), *Carmichael's manual of child psychology* (Vol. 2). New York: Wiley, 1970.

Goleman, D. Special abilities of the sexes: Do they begin in the brain? *Psychology Today,* November 1978, pp. 48–59.

Good, T., & Brophy, J. *Educational psychology: A realistic approach.* New York: Holt, Rinehart & Winston, 1977.

Gottesman, I., & Shields, J. A critical review of recent adoption, twin, and family studies of schizophrenia: Behavioral genetics perspective. *Schizophrenia Bulletin,* 1976, *2,* 360–400.

Gottfried, A. W. Intellectual consequences of perinatal anoxia. *Psychological Bulletin,* 1973, *80,* 231–242.

Gove, W., & Herb, T. Stress and mental illness among the young: A comparison of the sexes. *Social Forces,* 1974, *53,* 256–265.

Gove, W., & Tudor, S. Adult sex roles and mental illness. *American Journal of Sociology,* 1973, *80,* 812–835.

Graham, P., & Rutter, M. Adolescent disorders. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977.

Green, R. *Sexual identity conflict in children and adults.* Baltimore, Md.: Penguin Books, 1974.

Green, R. Atypical psychosexual development. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977.

Gunderson, J., Autry, J., & Mosher, L. Special report on schizophrenia, 1973. *Schizophrenia Bulletin,* 1974, *9,* 12–54.

Hall, C., & Lindzey, G. *Theories of personality* (2nd ed.). New York: Wiley, 1973.

Hanson, D., & Gottesman, I. The genetics, if any, of infantile autism and childhood schizophrenia. *Journal of Autism and Childhood Schizophrenia,* 1976, *6,* 209–234.

Hart, B. Gonadal androgen and sociosexual behavior of male mammals. *Psychological Bulletin,* 1974, *81,* 383–400.

Heincke, C. Learning disturbance in children. In B. Wolman (Ed.), *Manual of child psychopathology.* New York: McGraw-Hill, 1972.

Hersov, L. Adoption. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977. (a)

Hersov, L. Emotional disorders. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977. (b)

Hintgen, J., & Bryson, C. Recent developments in the study of early childhood psychoses: Infantile autism, childhood schizophrenia and related disorders. *Schizophrenia Bulletin,* 1972, *5,* 8–53.

Hoffman, L. W. Effects of maternal employment on the child—A review of the research. *Developmental Psychology,* 1974, *10,* 204–28.

Hoffman, L. W. Changes in family roles, socialization, and sex differences. *American Psychologist,* 1977, *32,* 644–657.

Johnson, D. Crosscultural perspectives on sex differences in reading. *The Reading Teacher,* 1976, *29,* 747–752.

Kagan, J., & Moss, H. *Birth to maturity.* New York: Wiley, 1962.

Kamin, L. Sex differences in susceptibility of IQ to environmental influence. *Child Development,* 1978, *49,* 517–518.

Kanner, L. Do behavior symptoms always indicate psychopathology? *Journal of Child Psychology and Psychiatry,* 1960, *1,* 17–25.

Kanner, L. *Childhood psychosis: Initial studies and new insights.* Washington, D.C.: V. H. Winston, 1973.

Keith, S., Gunderson, J., Reifman, A., Buschbaum, S., & Mosher, L. Special report: Schizophrenia, 1976. *Schizophrenia Bulletin,* 1976, *4,* 509–65.

Kessler, J. *Psychopathology of childhood.* Englewood Cliffs, N.J.: Prentice-Hall, 1966.

Kohlberg, L., LaCrosse, J., & Ricks, D. The predictability of adult mental health from childhood behavior. In B. Wolman (Ed.), *Manual of child psychopathology.* New York: McGraw-Hill, 1972.

L'Abate, L. Early infantile autism: A reply to Ward. *Psychological Bulletin,* 1972, *77,* 49–51.

Lefkowitz, M., & Burton, N. Childhood depression: A critique of the concept. *Psychological Bulletin,* 1978, *85,* 716–726.

Lehrke, R. A theory of X-linkage of major intellectual traits. *American Journal of Mental Deficiency,* 1972, *76,* 611–619.

Lehrke, R. Sex linkage: A biological basis for greater male variability in intelligence. In R. Osbourne, C. Noble, & N. Weyl (Eds.), *Human variation.* New York: Academic Press, 1978.

Maccoby, E. Sex differentiation during childhood. JSAS *Catalog of Selected Documents in Psychology,* 1976, *6,* 97. (Ms. No. 1339)

Maccoby, E., & Jacklin, C. *The psychology of sex differences.* Stanford, Calif.: Stanford University Press, 1974.

McClearn, G., & DeFries, J. *Introduction to behavioral genetics.* San Francisco: Freeman, 1973.

Mead, M. *Male and female.* New York: Morrow, 1949.

Miller, R. Childhood schizophrenia: A review of

selected literature. *International Journal of Mental Health,* 1974, *3,* 3–46.

Mischel, W. A social-learning view of sex differences in behavior. In E. Maccoby (Ed.), *The development of sex differences.* Stanford, Calif.: Stanford University Press, 1966.

Money, J. Intersexual and transsexual behavior and syndromes. In S. Arieti (Ed.), *American handbook of psychiatry* (Vol. 3). New York: Basic Books, 1974.

Money, J., & Ehrhardt, A. *Man and woman, boy and girl.* Baltimore, Md.: Johns Hopkins University Press, 1972.

Mosher, L., Gunderson, J., & Buschbaum, S. Special report: Schizophrenia, 1972. *Schizophrenia Bulletin,* 1972, *7,* 12–52.

Moss, H. Early sex differences and mother-infant interaction. In R. Friedman, R. Richard, & R. Vande Wiele (Eds.), *Sex differences in behavior.* New York: Wiley, 1974.

Moyer, K. The physiology of violence. *Psychology Today,* July 1973, pp. 35–38.

Musa, K., & Roach, M. Adolescent appearance and self-concept. *Adolescence,* 1973, *8,* 385–393.

Mussen, P. Early sex-role development. In D. Goslin (Ed.), *Handbook of socialization theory and research.* Chicago: Rand McNally, 1969.

Mussen, P., Conger, J., & Kagan, J. *Child development and personality.* New York: Harper & Row, 1974.

Ornitz, E., & Ritvo, E. The syndrome of autism: A critical review. *American Journal of Psychiatry,* 1976, *133,* 609–622.

Parlee, M. The premenstrual syndrome. *Psychological Bulletin,* 1973, *80,* 454–465.

Patterson, G., Littman, R., & Bricker, W. Assertive behavior in children: A step toward a theory of aggression. *Monographs of the Society for Research in Child Development,* 1967, *32*(4, Serial No. 113).

Pauly, I. B. Female transsexualism: I. *Archives of Sexual Behavior,* 1974, *3,* 487–507.

Peskin, H. Multiple prediction of adult psychological health from preadolescent and adolescent behavior. *Journal of Consulting and Clinical Psychology,* 1972, *38,* 155–160.

Phillips, S., King, S., & DuBois, L. Spontaneous activities of female versus male newborns. *Child Development,* 1978, *49,* 590–597.

Quadagno, D., Briscoe, R., & Quadagno, J. Effect of perinatal gonadal hormones on selected nonsexual behavior patterns: A critical assessment of the nonhuman and human literature. *Psychological Bulletin,* 1977, *84,* 62–80.

Reinisch, J., & Karow, W. Prenatal exposure to synthetic progestins and estrogens: Effect on human development. *Archives of Sexual Behavior,* 1977, *6,* 257–288.

Richman, N., Stevenson, J., & Graham, J. Prevalence of behavior problems in 3 year-old children: An epidemiological study in a London borough. *Journal of Child Psychology and Psychiatry,* 1975, *16,* 277–287.

Rimland, B. Comment on Ward's "Early infantile autism." *Psychological Bulletin,* 1972, *4,* 52–53.

Rosenthal, D. *Genetic theory and abnormal behavior.* New York: McGraw-Hill, 1970.

Rourke, B. P. Brain–behavior relationships in children with learning disabilities: A research program. *American Psychologist,* 1975, *30,* 911–920.

Ruble, D. Premenstrual symptoms: A reinterpretation. *Science,* 1977, *197,* 291–292.

Rutter, M. Concepts of autism: A review of research. *Journal of Child Psychology and Psychiatry,* 1968, *9,* 1–25.

Rutter, M. Sex differences in children's responses to family stress. In E. Anthony & C. Koupernki (Eds.), *The child in his family.* New York: Wiley, 1970.

Rutter, M. *Maternal deprivation reassessed.* Harmondsworth, England: Penguin Books, 1972. (a)

Rutter, M. Relationships between child and adult psychiatric disorders. *Acta Psychiatrica Scandinavica,* 1972, *48,* 3–21. (b)

Rutter, M. The development of infantile autism. *Psychological Medicine,* 1974, *4,* 147–163.

Rutter, M. Classification. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977. (a)

Rutter, M. Individual differences. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977. (b)

Rutter, M. Infantile autism and other child psychoses. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* London: Blackwell Scientific, 1977. (c)

Rutter, M. Diagnosis and definition of childhood autism. *Journal of Autism and Childhood Schizophrenia,* 1978, *8,* 139–161.

Rutter, M., Graham, P., Chadwick, O., & Yule, W. Adolescent turmoil: Fact or fiction? *Journal of Child Psychology and Psychiatry,* 1976, *17,* 35–56.

Rutter, M., & Yule, W. Reading difficulties. In M. Rutter & L. Hersov (Eds.), *Child psychiatry.* Oxford, England: Blackwell Scientific, 1977.

Schopler, E. National Society for Autistic Children: Definition of the syndrome of autism. *Journal of Autism and Childhood Schizophrenia,* 1978, *8,* 162–169.

Sears, R., Maccoby, E., & Levin, H. *Patterns of child rearing.* Evanston, Ill.: Row, Peterson, 1957.

Serbin, L., & O'Leary, D. How nursery schools teach girls to shut up. *Psychology Today,* December 1975, pp. 57–58.

Shah, S., & Roth, L. Biological and psychophysiological factors in criminality. In D. Glass (Ed.), *Handbook of criminality.* Chicago: Rand McNally, 1974.

Shepard, M., Oppenheim, B., & Mitchell, S. Childhood behavior disorders and the child guidance clinic: An epidemiological study. *Journal of Child Psychology and Psychiatry,* 1966, *7,* 39–52.

Sherman, J. Sex-related cognitive differences: An

essay on theory and evidence. Springfield, Ill.: Charles C Thomas, 1978.

Shields, S. A. Functionalism, Darwinism, and the psychology of women: A study in social myth. *American Psychologist*, 1975, *30*, 739–754.

Shinn, M. Father absence and children's cognitive development. *Psychological Bulletin*, 1978, *85*, 295–324.

Sorenson, R. *Adolescent sexuality in contemporary America: Personal values and sexual behavior ages 13 to 19*. New York: World, 1973.

Sullivan, H. *The interpersonal theory of psychiatry*. New York: Norton, 1953.

Tanner, J. Physical growth. In P. Mussen (Ed.), *Carmichael's manual of child psychology* (Vol. 1). New York: Wiley, 1970.

Tanner, J. Fetus into man: Physical growth from conception to maturity. Cambridge, Mass.: Harvard University Press, 1978.

Terman, L., & Tyler, L. Psychological sex differences. In L. Carmichael (Ed.), *Manual of child psychology*. New York: Wiley, 1954.

Toolan, J. Depression in children and adolescents. *American Journal of Orthopsychiatry*, 1962, *32*, 404–414.

Tresemer, D. Fear of success: Popular but unproven. *Psychology Today*, March 1974, pp. 82–85.

Vandenberg, B. Play and development from an ethological perspective. *American Psychologist*, 1978, *33*, 724–738.

Vellutino, F. Alternate conceptualizations of dyslexia: Evidence in support of a verbal-deficit hypothesis. *Harvard Educational Review*, 1977, *47*, 334–354.

Waber, D. Biological substrates of field dependence: Implications of the sex difference. *Psychological Bulletin*, 1977, *84*, 1076–1078.

Ward, A. Early infantile autism. *Psychological Bulletin*, 1970, *73*, 350–362.

Waters, E. The reliability and stability of individual differences in infant-mother attachment. *Child Development*, 1978, *49*, 483–494.

Watt, N., & Lubensky, A. Childhood roots of schizophrenia. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 363–375.

Watt, N., Stolorow, R., Lubensky, A., & McClelland, D. School adjustment and behavior of children hospitalized for schizophrenia as adults. *American Journal of Orthopsychiatry*, 1970, *40*, 647–657.

Willerman, L. *The psychology of individual and group differences*. San Francisco: Freeman, 1979.

Woerner, M., Pollack, M., Rogalski, C., Pollack, Y., & Klein, D. A comparison of the school records of personality disorders, schizophrenics, and their sibs. In R. Roff, L. Robins, & M. Pollack (Eds.), *Life history research in psychopathology* (Vol. 2). Minneapolis: University of Minnesota Press, 1972.

Wolff, S. Nondelinquent disturbances of conduct. In M. Rutter & L. Hersov (Eds.), *Child psychiatry*. Oxford, England: Blackwell Scientific, 1977.

Wolkind, S., & Rutter, M. Children who have been "in care"—an epidemiological study. *Journal of Child Psychology and Psychiatry*, 1973, *14*, 97–105.

Zuckerman, M., & Wheeler, L. To dispel fantasies about the fantasy-based measure of fear of success. *Psychological Bulletin*, 1975, *82*, 932–946.

Received January 27, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 3, 596–615

Heredity Versus Environment: An Integrative Analysis

Neil Gourlay
Department of Education
University of the Witwatersrand, Johannesburg, South Africa

Jencks's method of analysis of the heredity–environment data is presented, but with the important modification that cognizance is taken of the principle of genetic variation with age, that is, that the genotypic value (*G*) varies with age. As a result of this modification and a more critical examination of the original data on IQ, a solution is obtained that agrees with Jencks's figure for covariance, but supports the Burt–Jensen emphasis on heredity in that it assigns 75% of the remaining variance to heredity and only 25% to environment. The present study can be regarded as integrative in that (a) it eliminates most of the discrepancies in the field and (b) it uses Jencks's approach, albeit modified, to produce what is essentially Burt's result.

Of the many attempts at solving the problem of IQ heritability, probably the best known is that of the late Sir Cyril Burt. His claim that 75%–80% of IQ variance could be attributed to hereditary differences and only 20%–25% to environmental differences became general knowledge in the United Kingdom, with the publication of the article, "The Multifactorial Theory of Inheritance and its Application to Intelligence" (Burt & Howard, 1956). In the United States, however, Burt's work did not receive any marked degree of recognition until Jensen (1969) wrote in support of Burt's findings. Kamin (1974) aroused further interest in Burt 5 years later, but in this case it was the voice of criticism, directed mainly at Burt's handling of his data.

The solution that Jencks et al. (1972) offered in Appendix A of their book *Inequality* might be regarded as one of the strongest alternatives to that of Burt and Jensen. According to this solution, hereditary differences account for about 45%–50% of the total IQ variance and environmental differences for 30%–35%. The remaining variance (about 18%–19% of the total) is attributed to heredity–environment covariance, the concept that parents of good heredity not only pass on their superior genes to their children but also tend to provide them with superior intellectual environments.

In their 1956 article, Burt and Howard gave a figure of about 10% for covariance. But this was obtained for IQ group test data, and since they were primarily looking for data amenable to the application of Fisher's genetic theory (Fisher, 1918), they favored adjusted IQ assessments in which the attempt was made to eliminate systematic (i.e., between families) environmental effects. The result was the reduction of the covariance term virtually to zero. (It was this adjustment of IQs that became a main object of Kamin's criticism.)

Besides the efforts of Burt and Howard and Jencks et al., mention can be made of other approaches, such as multiple abstract variance analysis (MAVA; Cattell, 1960). Of particular importance is the work of two schools of biometrical genetics: the Birmingham school (Fulker, 1974; Jinks & Eaves, 1974; Jinks & Fulker, 1970) and the Hawaiian school (Rao, Morton, & Yee, 1974, 1976). Both these schools represent developments of the classical genetic theory laid down by Fisher (1918). The Birmingham school was influenced also by Mather (1949) and the Hawaiian school by Wright (1931). An important contribution of Wright was his development of path analysis, a technique that is used both by the Hawaiian school and by Jencks.

Requests for reprints should be sent to Neil Gourlay, who is now at the Department of Psychology, New University of Ulster, Coleraine, Londonderry, Northern Ireland BT52 1SA.

The essential difference between the two schools is their treatment of *dominance*, the concept that originated with Mendel, namely, that genes do not necessarily combine in a simple additive fashion at any one locus but can interact, thus adding to the genetic variance. (There is also the possibility of interaction between genes at different loci—what is known as *epistasis*.) Dominance is an important feature of the Birmingham school's approach to IQ heritability analysis. The Hawaiian school, on the other hand, chooses to exclude dominance from their genetic model, the argument being that it is more important to assess environmental effects (Rao et al., 1976).

Another major difference between the two schools lies in their estimates for covariance. The Hawaiian school, using Jencks et al.'s (1972) American data and a path model, have obtained an estimate on the order of 10%. The Birmingham school, using Burt's (1966) data, have found no evidence for covariance. This is to be expected, since Burt's data derive from adjusted assessments. But even in the case of Jencks et al.'s data, they have obtained the same zero result (Jinks & Eaves, 1974). This failure to detect covariance in Jencks et al.'s data throws considerable doubt on the power they claim for their method of weighted least squares (see below).

The Hawaiian school (Rao et al., 1976) have obtained the remarkable result that IQ heritability (i.e., the proportion of the total IQ variance to be attributed to hereditary factors) is very much greater for children than for adults: 67% compared with 21%. As I have shown in a recent publication (Gourlay, 1978), this result has practical (or empirical) implications that are very difficult to accept. For example, it follows that whereas the correlation between siblings (as children) is .53, the correlation between siblings (as adults) is .74. The difficulty is almost certainly due to the omission of dominance from the genetic model.

A feature of both schools is their use of high-powered statistical techniques—maximum likelihood and weighted least squares—in which all the data under consideration are combined in the one analysis with due weight being given to each item. Both these methods provide not only standard errors for the parameter estimates obtained but also tests of goodness of fit of the models employed.

On the face of it, the statistical methods used by the schools compare very favorably with the piecemeal type of analysis of Jencks, Burt, and others. On the other hand, the piecemeal approach permits the application of corrections for various error factors that affect the raw data (e.g., the selective placement factor in the case of foster-child studies). By not allowing for such factors, the parameter estimates of the schools are given a precision that can be quite spurious. Thus, applying weighted least squares to Jencks et al.'s data, Jinks and Eaves (1974) obtained an estimate of $.29 \pm .02$ for the environmental variance between families. The figure is almost certainly inflated by selective placement, for which Jinks and Eaves made no allowance. It is also significant that the application of their method to Jencks et al.'s (1972) data gave no evidence of covariance; apparently the method is just not sensitive enough to reveal the covariance that is undoubtedly there.

The present study does not use the methods of maximum likelihood or weighted least squares, but follows the general approach of Jencks. As justification, it can be argued (a) that by correcting or adjusting data for all relevant factors before the final analysis, one is more likely to obtain a solution that is consistent with all the known facts and (b) that this is preferable to the use of high-powered methods, which yield standard errors and tests of goodness of fit but produce parameter estimates that are obviously unacceptable empirically.

Although following the general approach of Jencks, the present study contains a number of important modifications, in one or two cases amounting to actual corrections:

1. A more critical examination is made of the data that derive from the four classical foster-child studies: Burks (1928), Freeman, Holzinger, & Mitchell (1928), Leahy (1935), and Skodak and Skeels (1949). This examination has the effect of largely removing the discrepancy that Jencks et al. found between the results obtained from the analysis of foster-child-study data and kinship correlations.

2. A new principle is introduced that has not so far appeared in heritability analysis, namely, that the genotypic value (G) is a function of age as well as ability. Its introduction removes two major difficulties that Jencks et al. encountered: (a) their failure to reconcile an item of the Skodak and Skeels data with their basic analysis (in the end, they dismissed the Skodak and Skeels study as "deviant" [Jencks et al., 1972, pp. 281–283]) and (b) the problem of explaining the large difference that exists between the correlations for siblings and fraternal twins (about 11%–12% of the total IQ variance). The general effect of introducing the new principle into heritability analysis is to increase the estimate of the hereditary component and to decrease the estimate of the environmental component.

3. In applying their basic path model (Jencks et al., 1972, pp. 279–281) Jencks et al. appeared to assume that the path coefficients from the child's G and E (environmental value) to his IQ were the same for the foster child as for the child living with his parents. This is an oversight and is corrected in the present analysis.

4. Jencks et al. have been criticized for not making their application of Fisher's (1918) genetic theory more explicit. Thus Jinks and Eaves (1974) felt that a weakness of their approach was their failure to deal systematically with dominance. Also, in one of their own notes (Jencks et al., 1972, p. 317) there is a cryptic reference to a comment by Crow that charges them with ignoring the effects of dominance. The present study attempts to set out more clearly Jencks et al.'s use of genetic theory. In particular, it corrects their formula for the genotypic correlation between fraternal twins, in which they confuse the genotypic correlation between spouses with the correlation between their additive deviations.

The outcome of the new analysis is a solution that appears to be consistent with all the data relevant to IQ heritability analysis. It supports Jencks in his estimate for covariance, but is closer to Burt and Jensen in its estimate of the relative importance of hereditary and environmental differences.

Theoretical Basis of Analysis

The Basic Model

IQ heritability analysis invariably starts with a simple additive model:

$$\textit{measured intelligence}\ (\text{IQ}) = G + E, \quad (1)$$

where in genetic terminology, IQ is the phenotypic, G the genotypic, and E the environmental value. The corresponding variance equation is

$$\text{var (IQ)} = \text{var}\ (G) + \text{var}\ (E) + 2\ \text{cov}\ (GE), \quad (2)$$

where cov (GE) is the covariance between G and E.

If one standardizes all three variables (i.e., with $M = 0$ and $SD = 1$), then, applying Jencks et al.'s notation, one can write

$$Q = hG + eE, \quad (3)$$

where Q represents IQ (standardized). The corresponding variance equation is now

$$1 = h^2 + e^2 + 2hes, \quad (4)$$

where h^2 and e^2 represent the fractions of the total IQ variance due to hereditary and environmental differences, respectively, and $2hes$ represents the fraction due to covariance, s being the correlation between G and E.

Genetic–environmental interaction. Some critics (e.g., Layzer, 1974) would insist that Equations 1 and 3 should also contain a term that represents genetic–environmental interaction (in the statistical sense). This would also mean additional terms in Equations 2 and 4. However, the most likely form of the interaction is a product term in G and E, and it can be shown that for such a term the variance is negligible. It follows that for the purpose of heritability analysis, it is fairly safe to omit genetic–environmental interaction from the basic model.

The age factor. Heritability analysis in the case of plants and animals (excluding humans) usually involves variables that characterize the mature organism, for example, size and quality of fruit or crop, milk and beef yields, and so on. The factor of age does not normally enter into the analysis. Consequently, it is generally assumed that (a) for a given population, h and e are fixed (i.e., heritability is a constant for the population) and (b) for any member of the population, G is fixed.

The same assumptions are to be found in the field of IQ heritability analysis (there is one exception). Yet, it is obvious that IQ herita-

bility analysis is different from plant and animal analysis in that since much of the data concerns children with ages ranging from 4 to 17 years, one is no longer dealing with a homogeneous population of mature individuals. Consequently, there is a distinct possibility that G varies with age and that even when heritability analysis is confined to children of the same age (x), the values of h and e also vary with x. It appears, therefore, that to allow for the age factor, Equation 3 should be written in the form

$$Q_x = h_x G_x + e_x E_x, \quad (5)$$

where x indicates that h, e, and G as well as Q and E can vary with age.

Following Wright (1931), the Hawaiian school of biometrical genetics have gone some way toward recognizing the importance of an age factor. They allow for the possibility that h and e can vary with age, but still make the standard assumption that G for the individual is fixed. As a result of their analysis, they give two values for h^2 and e^2: the values for a child population—a sort of average over all the ages involved—and the values for an adult population (Rao et al., 1976). The values are $h^2 = .670$ and $e^2 = [.094$ (common environment) $+ .135$ (residual)$]$ for the child population and $h^2 = .211$ and $e^2 = [.506$ (common environment) $+ .151$ (residual)$]$ for the adult, where *common environment* means common family environment. The estimates for the covariance ($2hes$) are .101 and .132, respectively.

As indicated earlier, these results have some rather strange empirical implications. In particular, it follows from the Hawaiian estimates that the correlation between the IQs of the adopted child and the natural mother is .26; but when the adopted child becomes an adult, the correlation falls to .15. This deduction is quite at variance with the Skodak and Skeels (1949) data, in which the correlation between adopted child and natural mother increases from about .28 to about .40 as the (average) age for adopted child moves from 4 years 3 months to 13 years 6 months.

The most likely explanation of the oddness of the Hawaiian results is the omission of dominance from their genetic model. But there is another important aspect to consider. Ignoring selective placement, it follows from Equation 5 that the IQ correlation between adopted child (AC) and natural mother (NM) is given approximately by

$$r(Q_{\mathrm{AC}x}, Q_{\mathrm{NM}}) = h_x r(G_{\mathrm{AC}x}, Q_{\mathrm{NM}}).$$

Consequently, if G_{AC} does not vary with x, $r(G_{\mathrm{AC}}, Q_{\mathrm{NM}})$ will be a constant, and the correlation between $Q_{\mathrm{AC}x}$ and Q_{NM} will vary directly with h_x. Thus if h_x decreases with x (the Hawaiian case), the correlation $r(Q_{\mathrm{AC}x}, Q_{\mathrm{NM}})$ will also decrease with x. On the other hand, if the correlation between the two IQs increases with x, as is suggested by the Skodak and Skeels data and as one might expect on a priori grounds, then either h_x must increase with x or $r(G_{\mathrm{AC}x}, Q_{\mathrm{NM}})$ must increase with x (or both). The first alternative is not likely. So it appears that the second alternative must hold, that is, the correlation between $G_{\mathrm{AC}x}$ and Q_{NM} will increase with x. It follows that G_{AC} (or simply G) varies with age x. This principle, that G varies with age, is an important feature of the analysis provided in this study. It is discussed more fully in the next section.

There remains the question of whether h and e also vary with age. Most likely there is some variation of h and e with age, and it is very probable that it takes the form of a decreasing h and an increasing e. However, for the purposes of the present study, it is assumed that no serious error will be incurred if, for ages 5 years and onwards, one takes h and e as constant and therefore also the covariance ($2hes$). The basic model for the study is therefore

$$Q_x = hG_x + eE_x, \quad (6)$$

where Q_x, G_x, and E_x vary with age x.

The Genotypic Value (G) *as a Function of Age and Ability*

Probably the main reason why past investigators have universally accepted a fixed G is that the genotype is fixed at conception. However, it must also be remembered that genetic development is a program in time that unfolds as the organism interacts with its environment. Different genetic factors operate at different points in time; moreover, the time schedule varies from individual to individual. It follows therefore that G must vary with age.

In the case of an intelligence test, such as the Stanford-Binet, this should be easily accepted, since it is well-known that the factor

content at any one age is different from that at another. But even in the case of a unidimensional variate such as height, the Gs for any two ages will not be the same (except, of course, at maturity). Obviously, in the case of different abilities, for example, the ability to read and the ability to do an intelligence test, it follows a fortiori that the Gs for one ability will not be the same as those for the other, even at the same point in time.

How does G for an IQ test such as the Stanford-Binet vary with age? Actually the pattern of variation is very similar to that for IQ itself. If the correlation between IQ at age 16 or 17 and IQ at an earlier age x is plotted against x, the curve obtained is a rising curve that levels out to the value of one at $x = 16$ or 17 (cf. Jensen, 1969, p. 18). In the case of G, the corresponding curve is similar in form and lies below the IQ curve, except at age 16 or 17 where the two curves come together. For the moment I simply demonstrate this result; later I derive it from the Skodak and Skeels data. (See Figure 1 on p. 612.) The demonstration depends on taking the results for a typical heritability analysis, say, those of Jencks et al. (1972), and assuming that h, e, and s are constant with age (see above).

According to Jencks, an acceptable analysis of the total IQ variance would be the following: heredity, 46.5%; environment, 35% (20% between families and 15% within a family); and covariance, 18.5%. Equationwise, Jencks et al.'s solution can be expressed

$$Q = .682G + .447EF + .387ER, \qquad (7)$$

where EF and ER are the between-families and within-family environmental measures and all four variates Q, G, EF, and ER have been standardized ($M = 0; SD = 1$).

Consider now a particular value of x, say, $x = 5$. According to Jensen (1969, p. 16), the correlation between IQs at ages 5 and 16 (or 17) is about .70, when correction is made for attenuation. It follows from Equation 7 that

$$.70 = .465r(G5, G16) + .20 + .15r(ER5, ER16) + .185,$$

where $G5$ represents the value of G at age 5, and so on. The correlation $r(ER5, ER16)$ cannot be zero. One can easily imagine that at least a quarter of the within-family environmental variance is accounted for by age 5. Let me therefore take .50 as the value for this correlation, recognizing that it could well be an underestimate. Substituting in Equation 6 one finds that $r(G5, G16) = .24/.465 = .516$. Obviously, the value of G at age 5 is considerably different from the value of G at age 16. Furthermore, it follows that the curve for $r(Gx, G16)$ lies below the curve $r(Qx, Q16)$.

The above has been worked out on the basis of Jencks et al.'s solution that did not apply the principle that G varies with age. Naturally, it would be just as easy to carry out the demonstration using a solution that incorporated the principle.

Implications of the variability of G *with age.* The recognition of the principle of the variability of G with age has two immediate consequences. First, any definition of G must specify not only the population and the IQ test for which G is being defined but also the age group. Following Falconer (1967, p. 113), one might define the genotypic value for a given IQ test at age x as follows:

If one would replicate the genotype in a number of individuals and measure them for IQ at age x, after they lived in environmental conditions normal for the population, their mean environmental deviation would be zero, and their mean phenotypic value (i.e., IQ) would consequently be equal to the IQ genotypic value of that particular genotype at age x.

Second, the basic model must show that G as well as E and Q varies with age. This, of course, has already been done (see Equation 6). There are several other important consequences:

1. The variation in the correlation between IQ at 16 years of age and IQ at earlier ages has often been quoted as evidence of the inconstancy of IQ, that is, that IQ varies considerably as a result of environmental change. The introduction of the principle that the genotypic value varies with age provides a different explanation for the inconstancy, namely, that most of the variation is of a genetic character. Even on the basis of the simple demonstration presented earlier, it appears that environmental change accounts for no more than a quarter of the change in the IQ correlation with age.

There is an obvious parallel with Piaget's theory of stages in cognitive development. Piaget's theory, although interactionist, is essentially genetic in character—certainly implicitly so—and it is possible to argue that the principle of the variability of G with age is a translation of Piaget into statistical terms.

2. The correlation between the genotypic values of siblings is not independent of the ages of the siblings—as is generally assumed—but varies with the age interval between them. The greater the age interval, the smaller the correlation. In fact, if one considers children of age 16 and their siblings at earlier ages x, the correlation between the Gs of the sibling pairs (one at age 16 and one at age x) varies with x in the same manner as $r(Gx, G16)$, except that the maximum value (at $x = 16$) is on the order of .50.

It follows that the correlation between the IQs of ordinary siblings is depressed vis-à-vis the correlation for fraternal twins—not just because of greater environmental differences but also because of the greater differences in their Gs. This, of course, clears up the problem, which obviously worried Jencks, of how to explain the large difference between the correlations for siblings (different ages) and those for fraternal twins (same age): .59 versus .70 (Jencks et al., 1972, pp. 289; 290).

3. In the same way, the correlations between the Gs for child and parent vary with the age of the child. Skodak and Skeel's (1949) study is unique in that it provides confirmatory empirical evidence. The result also explains Burt's (1966) finding that the correlation between parent (adult) and child is .50, whereas the correlation for parent (as child) and child is .56 (cf. Jensen, 1969, p. 49; Kamin, 1974, pp. 94; 95).

4. It should now be apparent that the principle of G variability with age has important implications for heritability analysis. Previous analyses have been wrong in that they have ignored differences in age between parent and child and between siblings. Obviously, analyses must be carried out on data for which the Gs are at the same age level, and where the data are not in this form to begin with, they must be adjusted so that the Gs are comparable. Oddly enough, Burt, who has been subjected to so much criticism of late, largely avoided the problem of G variability by concentrating on twin data in which age differences did not operate.

The Genetic Model

It follows from the definition of G, as also from the basic model (Equation 6), that G includes all the underlying genetic factors. In the literature, there is some confusion on this point in that the term *genotypic value* is sometimes applied only to additive genetic effects (genetic interaction effects such as dominance are treated as phenotypic). However, for the purposes of this study, it must be made clear that G embraces not only additive effects but any other genetic effects one chooses to consider.

An examination of IQ heritability studies shows that only three main genetic effects are considered: (a) additive effects if mating were random, (b) additive effects due to assortative mating (the concept that mating is not random and that the additive effects of spouses are correlated), and (c) dominance (the concept that the genes at any one locus can interact and produce a nonadditive effect). Other genetic factors are possible, for example, epistasis (the concept of interaction between genes at different loci), but it is generally assumed that these effects are not important for IQ heritability analysis and can be ignored.

In restricting genetic factors to these three main effects, IQ heritability analysis follows the classical genetic theory of Fisher as laid down in his 1918 article. The present study, like other studies, does not attempt to advance on this theoretical basis.

Expressing the genetic model in equation form, one can write $G = A + D$, where A = additive effects (including assortative mating) and D = dominance effects. The corresponding variance equation is

$$V_G = V_A + V_D. \tag{8}$$

There is no covariance term, A and D being uncorrelated.

The additive variance V_A is often termed the genic variance and can be split into two subcomponents: (a) the variance due to additive effects if mating were random and (b) the additional additive variance due to assortative

mating. Burt (1975, p. 127) called these two components the additive variance (V_A) and the variance due to assortative mating (V_{AM}). But to avoid confusion, I call them both additive variance and denote them by $V_{A(R)}$ and $V_{A(AM)}$; that is, Equation 8 can be written

$$V_G = V_{A(R)} + V_{A(AM)} + V_D.$$

I now standardize G, A, and D (with $M = 0$ and $SD = 1$) and write

$$G = aA + dD, \tag{9}$$

where

$$1 = a^2 + d^2, \tag{10}$$

$$a^2 = V_A/(V_A + V_D),$$

and

$$d^2 = V_D/(V_A + V_D);$$

that is, a^2 and d^2 are the proportions of the genotypic variance to be attributed to additive effects and dominance, respectively. Also, it should be noted that the term a^2 is equivalent to c_2 in Fisher's (1918) analysis.

I now consider the relationship between the genetic measures for parents and offspring. In the discussion that follows, it is understood that the Gs, As, and Ds have been defined for the same age (or level) of development; that is, either children's values have to be defined for the adult level or parents' values have to be defined for the same age as the children. I assume the former. The Gs, As, and Ds for father, mother, and offspring are distinguished by subscripts F, M, and C, respectively.

Since the child gets half of his genes from each parent, it follows that

$$A_C = .50(A_F + A_M) + \textit{random term} \text{ (within family)}. \tag{11}$$

A similar equation can be written that expresses G_C in terms of G_F and G_M, namely,

$$G_C = g(G_F + G_M) + \textit{random term} \text{ (within family)}, \tag{12}$$

where the same symbol (g) is used as in Jencks et al. (1972).

The coefficient g will not be .50, since G includes dominance as well as additive effects. It is important that one obtain an expression for its magnitude.

As a first step, it might be noted that the correlation between G_C and the G of either parent (G_P) is

$$r(G_C, G_P) = g[1 + r(G_F, G_M)],$$

which can be written

$$r(G_C, G_P) = g(1 + p), \tag{13}$$

where $r(G_F, G_M) = p$ (Jencks et al.'s notation) $= \mu$ (Fisher's notation). (It might be noted that Fisher used the same symbol μ to denote both the correlation between the genotypic values and the correlation between the phenotypic values.)

The correlation $r(G_F, G_M)$ implies assortative mating. It is assumed as in Jencks et al. that G_F and G_M are correlated through the IQs of the parents. It follows that

$$p = r(G_F, G_M) = r(G_F, Q_F)r(Q_F, Q_M)r(Q_M, G_M).$$

But by Equation 3,

$$r(G_F, Q_F) = h + es = r(Q_M, G_M).$$

Hence

$$p = (h + es)^2 r(Q_F, Q_M), \tag{14}$$

which is Jencks et al.'s (1972, p. 273) result.

Similarly, following Fisher's genetic model, as did Jencks et al., A_F and A_M are correlated through G_F and G_M, where

$$r(A_F, A_M) = a^2 r(G_F, G_M) \tag{15}$$

(see Equation 9).

Denoting $r(A_F, A_M)$ by A as in Fisher (Jencks et al. made the mistake of omitting A from their analysis), one has $A = a^2 p = c_2\mu$ (Fisher's notation). (This equation is not to be confused with the other well-known equation of Fisher, $A = c_1c_2\mu$. As noted earlier, Fisher used μ in two senses. When μ denotes the correlation between the *genotypic* values of parents, $A = c_2\mu$. But when μ denotes the correlation between their *phenotypic* values, $A = c_1c_2\mu$. It would have been better if Fisher had used two symbols, say, μ_g and μ_p.)

Corresponding to Equation 15, one also has $r(D_F, D_M) = d^2p$ and $r(A_F, D_M) = adp = r(A_M, D_F)$. By means of these equations, one can derive another expression for $r(G_C, G_P)$

Table 1
Summary of Relationships Between Genetic Components in Fisher's (1918) and Jencks et al.'s (1972) Terminology

Component	Fisher	Jencks et al.
$r(G_F, G_M)$	μ	$p = (h + es)^2 r(Q_F, Q_M)$
$a^2 = V_A/(V_A + V_D)$	c_2	$2g$
$r(A_F, A_M)$	$A = c_2\mu$	$2gp$
$r(G_C, G_P)$	$.5c_2(1 + \mu)$	$g(1 + p)$
$r(G_1, G_2)$	$.25(1 + c_2 + 2c_2A)$	$.25 + .5g + 2g^2p$
$V_{A(R)}/(V_A + V_D)$	$c_2(1 - A)$	$2g(1 - 2gp)$
$V_{A(AM)}/(V_A + V_D)$	c_2A	$4g^2p$
$d^2 = V_D/(V_A + V_D)$	$1 - c_2$	$1 - 2g$

as follows:

$$r(G_C, G_P) = E[(aA_C + dD_C)(aA_P + dD_P)]$$
$$= E\{(aA_P + dD_P)[.50a(A_F + A_M) + dD_C + \textit{random term}]\}$$

(by Equation 11; where E = expected value). Assuming that the correlation between D_C and D_P is negligible, this becomes

$$r(G_C, G_P) = .50a^2(1 + A) + .50ad(adp),$$

which reduces to

$$r(G_C, G_P) = .50a^2(1 + p). \tag{16}$$

Comparing Equations 13 and 16, it is seen that $g = .50a^2 = .50c_2$ (Fisher's notation); that is, c_2(Fisher) $= 2g$(Jencks) $= a^2$.

Since it is required later, I also derive the correlation between the genotypic values for fraternal twins (or siblings at the same age). Thus, one can write

$$r(G_1, G_2) = E[(aA_1 + dD_1)(aA_2 + dD_2)]$$
$$= E\{[.50a(A_F + A_M) + dD_1 + \textit{random term}] \times [.50a(A_F + A_M) + dD_2 + \textit{random term}]\}.$$

It must now be noted that whereas the Ds for offspring are uncorrelated with the As and Ds for parents, the Ds for siblings are correlated. I assume the standard figure for the correlation, namely, .25 (cf. Falconer, 1967, p. 157). It follows that

$$r(G_1, G_2) = .25a^2(2 + 2A) + .25d^2,$$

which, using Equation 10, reduces to

$$r(G_1, G_2) = .25 + .25a^2 + .50a^2A$$
$$= .25(1 + c_2 + 2c_2A) \text{ (Fisher)}$$

or

$$r(G_1, G_2) = .25 + .50g + 2g^2p \text{ (Jencks)}. \tag{17}$$

This equation corrects Jencks et al.'s mistake (1972, p. 303): They seem to have assumed that p, the correlation between parental Gs, is the same as the correlation between their additive deviations.

Table 1 summarizes a number of the above results and includes expressions for the components of the genetic variance. In each case, both Fisher's and Jencks's terminology are presented.

The Analysis

The general approach is that of Jencks et al. (1972). In other words, a solution is sought mainly through data on the correlations between parents and children, both natural and foster. The only other item of data of equal importance is the correlation for fraternal twins (of the same sex). The twins correlation has the advantage over the correlation for ordinary siblings in that because twins are of the same age, their Gs are comparable. This permits the immediate application of Fisher's (1918) genetic theory (see, in particular, Equation 17).

As in Jencks, all correlations are corrected for attenuation. Error terms do not therefore appear in the variance analysis.

Notation

The symbols NC and AC are used to denote natural (own) and adopted child; NP, NF, and NM and AP, AF, and AM are used to denote natural and adopting parent, father, and mother, respectively. The symbol Q is used for IQ (standardized); but in the case of correlations involving IQs, it is normally omitted from the notation, except where there could be confusion. Thus, $r(\mathrm{F}, \mathrm{M})$ denotes the correlation between the IQs of parents (corrected for attenuation) and $r(\mathrm{AC}x, \mathrm{NM})$ denotes the correlation between the IQs of adopted child at age x and natural mother. On the other hand, a correlation such as $r(G_{\mathrm{NC}x}, Q_{\mathrm{NM}})$ retains the Q in the notation.

Three expressions that occur frequently in the analysis are represented by special symbols. These are $\alpha = (h^2 + e^2)^{\frac{1}{2}}$, $Y_x = r(\mathrm{NC}x, \mathrm{NP})$, and $Z = r(\mathrm{AC}, \mathrm{AM}; \mathrm{NSP})$, where NSP indicates the correlation between AC and AM with no selective placement.

Basic Data

A large part of the data is taken from Jencks et al. (1972), who provide an excellent summary of American data related to IQ heritability analysis. The first two important items are (a) the correlation between the IQs of father and mother (the empirical measure of assortative mating), $r(\mathrm{F}, \mathrm{M}) = .57$, and (b) the correlation between the IQs of parent and own child living with parent.

Jencks et al.'s study gives an estimate of .55 for the latter correlation, but the statistic has to be qualified. One must remember that in accordance with the principle of the variability of G with age, $r(\mathrm{NC}, \mathrm{NP})$ will vary with the age level of the children from whose IQs the correlation was derived—The younger the children, the lower the correlation. It is assumed that the figure of .55 applies to children whose ages are about average for foster-child studies. I take this average as 9.3 years, the figure for the Leahy (1935) study. I therefore write

$$Y_{9.3} = r(\mathrm{NC}9.3, \mathrm{NP}) = .55. \qquad (18)$$

Also important is the correlation between the IQs of fraternal twins. It is assumed that the correlation is independent of the age level of the twins. Also, like Jencks, I take the figure provided by the Newman, Freeman, and Holzinger (1937) study; that is, I write

$$r(1, 2) = .692 = r_{\mathrm{IQDZT}} \quad \text{(Jencks et al.'s notation)}. \qquad (19)$$

This statistic is considerably larger than the .59 (corrected for attenuation) reported by Burt (1966). However, as Kamin (1974) pointed out, there are several large-scale studies that support the Newman et al. figure. Also, as has already been argued, the principle of the variability of G with age implies a figure rather higher than the .57 generally accepted for siblings. (One must remember also that the factor of greater environmental differences for siblings vis-à-vis twins further increases the difference between the two correlations.) Finally, Jencks et al. provide data from the four foster-child studies: Burks (1928), Freeman (1928), Leahy (1935), and Skodak and Skeels (1949).

Together the studies provide an estimate of $Z = r(\mathrm{AC}, \mathrm{AM}; \mathrm{NSP})$. The Skodak and Skeels study is of particular importance in that it furnishes data that are used to derive the variation of $r(G_{\mathrm{NC}x}, Q_{\mathrm{NM}})$ with age x and also of $r(Gx, G16)$ with x.

Basic Equations

The equations required for the analysis are set out below. Their derivation, together with underlying assumptions, is briefly indicated in the Appendix.

1. For children living with their natural parents,

$$Q_{\mathrm{NC}x} = hG_{\mathrm{NC}x} + eE_{\mathrm{NC}x}. \qquad (20)$$

This, of course, is the same as Equation 6.

2. For foster children with no selective placement,

$$Q_{\mathrm{AC}x} = (1/\alpha)(hG_{\mathrm{AC}x} + eE_{\mathrm{AC}x}), \qquad (21)$$

where $\alpha = (h^2 + e^2)^{\frac{1}{2}}$. Obviously, it is assumed that the genetic and environmental variances are the same for foster children as for own children or, at least, that the data can be adjusted in accordance with this assumption. The covariance for foster children is of course zero if there is no selective placement.

In the case of a moderate degree of selective placement (as for the Burks, Leahy, and Skodak and Skeels studies—Freeman is the exception), it can be shown that little error is involved in assuming that Equation 21 still holds (see Appendix).

3. For foster child and foster parent (no selective placement),

$$Z = r(\mathrm{AC}, \mathrm{AP}; \mathrm{NSP}) = (e/\alpha)r(E_{\mathrm{NC}}, Q_{\mathrm{NP}}). \quad (22)$$

4. For foster child and foster mother (SP refers to selective placements),

$$r(\mathrm{AC}x, \mathrm{AM}; \mathrm{SP}) = Z + (h/\alpha)r(G_{\mathrm{NC}x}, Q_{\mathrm{NM}})r(\mathrm{NM}, \mathrm{AM}); \quad (23)$$

that is, the correction for obtaining Z from

$$r(\mathrm{AC}x, \mathrm{AM}; \mathrm{SP}) = -(h/\alpha)r(G_{\mathrm{NC}x}, Q_{\mathrm{NM}})r(\mathrm{NM}, \mathrm{AM}). \quad (24)$$

5. For parent and own child,

$$Y_x = r(\mathrm{NC}x, \mathrm{NP}) = hr(G_{\mathrm{NC}x}, Q_{\mathrm{NP}}) + \alpha Z. \quad (25)$$

Assuming that for $x = 16$, the Gs for children and parents are comparable, one can write

$$hr(G_{\mathrm{NC}16}, Q_{\mathrm{NP}}) = gh(h + es)[1 + r(\mathrm{F}, \mathrm{M})], \quad (26)$$

where g is defined as earlier (see Equation 12).

6. For covariance,

$$2hes = \frac{4\alpha Z}{1 + r(\mathrm{F}, \mathrm{M})}(Y_{16} - \alpha Z). \quad (27)$$

7. For foster child and natural mother,

$$r(\mathrm{AC}x, \mathrm{NM}; \mathrm{NSP}) = r(\mathrm{AC}x, \mathrm{NM}; \mathrm{SP}) - Zr(\mathrm{NM}, \mathrm{AM}), \quad (28)$$

where the second term on the right-hand side is the correction for selective placement. Also,

$$r(\mathrm{AC}x, \mathrm{NM}; \mathrm{NSP}) = (h/\alpha)r(G_{\mathrm{NC}x}, Q_{\mathrm{NM}}). \quad (29)$$

8. For fraternal twins (of the same sex),

$$r(1, 2) = h^2 r(G_1, G_2) + 2hes + e^2 r(E_1, E_2), \quad (30)$$

where it is assumed that

$$s = r(G_1, E_1) = r(G_2, E_2) = r(G_1, E_2) = r(G_2, E_1).$$

For identical twins living together, it is generally accepted that the environmental variance within a pair of twins is .03 (3%). According to Jencks et al. (1972, p. 308), the environmental variance within a fraternal twin pair is not likely to be much more than .03. But some critics would regard this as a rather tenuous assumption. Thus, Kamin (1974, p. 99), referring to data from same-sex and opposite-sex fraternal twins, argued that the environmental variance for fraternals must be considerably higher than that for identicals. However, there is evidence that the IQ variance for boys is greater than that for girls (cf. Publications of the Scottish Council, 1949), and it can be argued that the data to which Kamin refers are more easily explained on the basis of greater genetic variance for boys than for girls. This would mean, of course, that there should be a difference in IQ heritability between boys and girls and that the result of all "mixed" analyses, like that of the present study, is only an average for the sexes.

Nevertheless, for the moment, I treat the environmental variance within a fraternal twin pair as an unknown quantity and denote it by ω; that is, $e^2 r(E_1, E_2) = e^2 - \omega$. Equation (30) then becomes

$$r(1, 2) = h^2 r(G_1, G_2) + 2hes + e^2 - \omega. \quad (31)$$

The genetic correlation $r(G_1, G_2)$ is given by Equations 14 and 17. Substituting in Equation 31, one obtains

$$r(1, 2) = h^2[.25 + .50g + 2g^2(h + es)^2 r(\mathrm{F}, \mathrm{M})] + 2hes + e^2 - \omega. \quad (32)$$

Outline of Analysis

The first step is to obtain an estimate of $Z = r(\mathrm{AC}, \mathrm{AM}; \mathrm{NSP})$. The figure adopted was the average of four estimates, one from each of the four foster-child studies. Since this involved the application of corrections for selective placement, dependent for their value on the final solution, a reiterative method had to be employed for the analysis as a whole; that is, initially approximations to these corrections were used. Then when a solution was obtained that had been derived from these values, the corrections were revised. The

procedure was repeated until no further revision was required.

For simplicity of presentation, the reiterations are omitted from the account that follows, and only the final values for the corrections appear. Since $\alpha = (h^2 + e^2)^{\frac{1}{2}}$ is involved in these corrections, the final value for α is assumed in making this simplified presentation. It follows that in completing the analysis, it is necessary to show that the values finally obtained for h^2 and e^2 are consistent with the initially assumed value for α. The remaining steps of the analysis are as follows:

1. Values are found for $hr(G_{NC16}, Q_{NP})$ in Equation 26 and for Y_{16} in Equation 27. Details of this are presented later. A value for the covariance ($2hes$) follows immediately from Equation 27; also, Equation 26 can be expressed entirely in terms of the four unknowns, g, h, e, and s.
2. Equations 26, 27, 32, and 4 can then be solved so that values for g, h, e, and s are obtained for a range of values of ω (the environmental variance within a fraternal twin pair).
3. Finally, a suitable value for ω is determined from a consideration of the likely values for e^2 (the total environmental variance) and its between-families and within-family components. Data used in making this decision are (a) the difference in the correlations for fraternal twins and ordinary siblings and (b) the results of studies of separated identical twins. With the choice of a value for ω, the other unknowns (g, h, e, and s) are fixed. One can then write down the contributions of heredity, environment, and covariance to the total IQ variance, and, using the formulae in Table 1, one can derive the $V_{A(R)}$, $V_{A(AM)}$, and V_D components of the genetic variance.

The Correlation Z = r(*AC, AM; NSP*)

The Burks (*1928*) *study.* Burks's study yielded the following statistics: r(AC, AM) = .23 and r(AC, AF) = .09, both correlations corrected for attenuation.

The correlation r(AC, AF) is abnormally low and can be shown to be the cause of the spuriously high value for the between-families environmental variance (17.6%) that Burks obtained from her multiple regression analysis. I therefore deal only with r(AC, AM). In any case, selective placement appears to operate mainly through the mothers (natural and foster); also, statistics on selective placement are usually with respect to mothers. The correction for selective placement is given by Equation 24. Assuming a value of .16 for r(NM, AM) (cf. Jencks et al., 1972, p. 277) and applying the reiterative method, one obtains a value of −.062 for the correction (the final value of α = .902); that is, it appears that Z(Burks) = .23 − .062 = .168.

A further correction for restriction in range of family environment might be necessary, but it is not likely to be large. One must simply bear in mind that the figure of .168 could be on the low side.

The Leahy (*1935*) *study.* The values obtained for r(AC, AP) were as follows: r(AC, AM) = .20 and r(AC, AF) = .15, or .214 and .161 (approximate) when corrected for attenuation. In her study, Leahy did not correct for attenuation, but did apply a correction for restriction in range. (The standard deviation of the IQs for the foster children was only 12.5.) This correction gave her values of .24 and .19, respectively, and it was these values that Jencks et al. combined with the Burks and Freeman (1928) values to get their overall mean for r(AC, AP), uncorrected for attenuation. However, a closer examination of Leahy's study shows that the small value for the standard deviation of the foster children's IQs was due mainly to a restriction in the genetic range of the foster children; a correction for this must reduce the correlations and not increase them. There was, nevertheless, some restriction in the environmental range of the foster children, as is evidenced by the fact that the standard deviation of the environmental status score was 54.3 for foster children and 59.6 for controls.

When all the necessary corrections are made, it is found that the correlation between IQ_{AC} and IQ_{AM} amounts to about .18 (because of their lengthiness, these calculations are not reproduced here); that is, Z(Leahy) = .18.

The Freeman (*1928*) *study.* This is the largest of the foster-child studies, involving 401 foster children. Correlations yielded by this study are considerably higher than those obtained in the Burks and Leahy studies. Since the average age of adoption was 4 years 2

months, selective placement is an obvious explanation, but this was resisted strongly by Freeman. The values of r(AC, AP) obtained by Freeman (uncorrected for attenuation) were r(AC, AM) = .28 and r(AC, AF) = .34. Another important statistic in Freeman's study is $r(Q_{AC}, HR) = .48$, where HR is the home rating of the foster home.

In a reanalysis of Freeman's data, to be submitted for publication, the difficulty of selective placement was avoided by a comparatively simple technique. Instead of considering the correlation between the IQs of foster children and the *HR*s of the foster homes—a procedure that cannot avoid the factor of selective placement—I estimated the IQ gains of the foster *groups* as a result of being moved from natural to foster homes. Estimates were also derived for the mean *HR*s of the natural homes—Freeman gave only the *HR*s for the foster homes—and, as a result, a measure of the mean gain in IQ per unit increase in HR was obtained. This, of course, can easily be converted to a correlation. In this way, I obtained a revised measure for $r(Q_{AC}, HR)$, which, when corrected for attenuation, amounted to only .295.

The question arises, What is the corresponding value of r(AC, AM)? It is obviously something less. As a result of further calculation, I was able to show that this correlation must be about .20; that is, Z(Freeman) = .20.

The Skodak and Skeels (1949) study. This study does not provide values for r(AC, AP). But it has the special feature of providing values for r(AC, NM) at different ages of the foster child.

The foster children were tested on four occasions, at mean ages of 2 years 2 months, 4 years 3 months, 7 years 0 months, and 13 years 6 months. In the case of 63 children (40 girls and 23 boys), the IQs were also available for the natural mothers. In all cases, the Stanford-Binet scale was used. But at the last testing (M = 13 years 6 months) the children were also given the 1937 Terman-Merrill scale. This was done in view of the inaccuracy of the Stanford-Binet standardization at the older ages.

Skodak and Skeels provided a magnificent appendix to their study, and, using the data provided there, I was able to obtain the values of r(AC, NM) at the different ages, for both boys and girls and for the two together (see Table 2). Two of the values obtained for the total group differ slightly from the correlations provided by Skodak and Skeels (in particular, the Terman-Merrill value for 13 years 6 months).

Table 2
Correlations Between IQs of Foster Child (AC) and Natural Mother (NM) for Different Ages of the Foster Child

M age (*years–months*)	r(AC, NM) Girls	Boys	Total
2–2[a]	.202	−.248	.037 (.00)[b]
4–3	.299	.233	.275 (.28)
7–0	.363	.330	.348 (.35)
13–6	.520	.066	.381 (.38)
13–6[c]	.472	.292	.415 (.44)

Note. The correlations were derived from Skodak and Skeels's (1949) data and are uncorrected for attenuation; for girls, n = 40; for boys, n = 23.
[a] Both children's and mothers' IQs were derived from the Stanford-Binet scale.
[b] Correlations in parentheses are figures given by Skodak and Skeels (1949).
[c] Children's IQs were derived from the Terman-Merrill scale and mothers' IQs from the Stanford-Binet scale.

Three things should be noted:

1. The effect of selective placement, as Jencks et al. (1972, p. 282) pointed out, is relatively unimportant. The correction is $-Zr$(NM, AM) (see Equation 28), which for r(NM, AM) = .16 and Z = .20 works out at −.032 (approximate). (It is understood that in correcting for selective placement, the same correction has to be applied to all correlations in Table 2.)

2. From an examination of Table 2, it is seen that for both girls and the total, the correlations increase steadily with chronological age. When a curve is fitted to the data, one obtains a rising curve that flattens out at the later ages. In the case of the correlations for the boys, the same trend is discerned, but the figures are very erratic. Sampling error obviously comes to mind.

3. With the number of boys and girls amounting to only 40 and 23, respectively, the degree of sampling error can be large. If one thinks in terms of a rising curve, the effect is

to raise or lower the position of the curve but not to change its general shape. Later analysis appears to indicate that sampling error is positive for the girls and distinctly negative for the boys. Furthermore, partly as a result of its greater homogeneity and smaller size, the boys' group shows greater fluctuation in its correlations. Also, despite the general homogeneity of the boys' group, the IQs for one boy and his natural mother lie well outside the range of the others, and this, together with the faulty Stanford-Binet standardization, adds considerably to the erratic variations in the correlations for the group. These irregularities also show to some extent when the boys are combined with the girls and the correlations for the total group are obtained.

In view of these considerations, it is obvious that one has a certain amount of difficulty in deciding what one should take as the value of $r(\mathrm{AC, NM})$ for the four mean ages of the Skodak and Skeels study. Despite the obvious differences in the boys' and girls' groups, I choose the values for the total group; in the case of the oldest of the four ages, 13 years 6 months, where the value is critical, I somewhat arbitrarily choose the figure .415. This estimate has at least the advantage of being almost the same as the .41 assumed by Jencks et al. (1972, p. 282).

Thus, for the four Skodak and Skeels mean ages, the values of $r(\mathrm{AC, NM})$ are .037, .275, .348, and .415 or, correcting for attenuation, .040, .299, .378, and .451.

When I now apply the correction for selective placement (−.032), the values become .008, .267, .346, and .419. In addition to these values, one requires the values for ages 8.2 and 9.3, the respective mean ages of the Burks and Leahy studies and also the value for age 16. By simple interpolation and extrapolation, these values are found to be .364, .384, and .423 (approximate); that is, at age $x = 9.3$,

$$r(\mathrm{AC9.3, NM; NSP}) = .384,$$

and therefore by Equation 29,

$$hr(G_{\mathrm{NC9.3}}, Q_{\mathrm{NM}}) = .384\alpha = .346$$

(assuming that the final value for $\alpha = .902$).

It now follows from Equation 25 that $\alpha Z = .55 - .346 = .204$, whence $Z = .204/.902 = .226$; that is, $Z(\text{Skodak and Skeels}) = .226$.

The averaged result. One now has four estimates for $Z = r(\mathrm{AC, NM; NSP})$, namely, Burks, .168; Freeman (revised), .200; Leahy, .180; and Skodak and Skeels, .226. The agreement is remarkably good. The average of the four estimates is .194, but since the figure for the Burks study is probably on the low side, I take .20 as the final estimate; that is,

$$Z(\text{final estimate}) = .20. \qquad (33)$$

This estimate is somewhat less than that of Jencks et al., which appears to be about .23 on correction for selective placement. One of the main reasons for the difference is Jencks et al.'s acceptance of the abnormally high figures produced by the original Freeman study. Another contributing factor is the mistake made by Leahy (in correcting her correlations for restrictions in range). Also, it is seen that there is now no need to reject the Skodak and Skeels study as "deviant," as did Jencks.

Derivation of Values for $\mathrm{hr}(G_{NC16}, Q_{NM})$ *and* $Y_{16} = \mathrm{r}(NC16, NM)$

As was pointed out earlier in the outline of the analysis, values are required for these two expressions in order to simplify two of the four equations required for the final solution. In deriving the values, it is necessary to correct the Skodak and Skeels $r(\mathrm{AC, NM; NSP})$ correlations for sampling error. The procedure is the reverse of that followed in estimating Z for the Skodak and Skeels study.

Assuming $\alpha = .902$ and using the overall estimate of Z as given by Equation 33,

$$\alpha Z = .902 \times .20 = .180. \qquad (34)$$

From Equations 18 and 25, it follows that

$$hr(G_{\mathrm{NC9.3}}, \mathrm{NM}) = .550 - .180 = .370.$$

Substituting in Equation 29,

$$r(\mathrm{AC9.3, NM; NSP}) = .370/902 = .410.$$

Comparing this with the value of .384 provided by the Skodak and Skeels study, it is now seen that sampling error, for boys and girls together, is small and involves a correction of only + .026; that is, the final values of $r(\mathrm{AC, NM; NSP})$ for the four ages involved in the Skodak and Skeels study are .034, .293,

.372, and .445. For ages 8.2 and 9.3, the interpolated values are .390 and .410. Last, for age 16 the extrapolated value becomes .449.

By Equation 29, it follows that

$$hr(G_{\mathrm{NC16}}, Q_{\mathrm{NM}}) = .449\alpha = .405. \quad (35)$$

Also, by Equation 26,

$$gh(h + es)[1 + r(\mathrm{F, M})] = .405$$

or, substituting for $r(\mathrm{F, M})$,

$$1.57gh(h + es) = .405. \quad (36)$$

Last, by Equations 25, 34, and 35,

$$V_{16} = r(\mathrm{NC16, NM}) = .405 + .180 = .585. \quad (37)$$

*The Covariance 2*hes

The relevant equation is 27. By Equations 34 and 37, it follows that

$$2hes = .186. \quad (38)$$

Also, $\alpha = (h^2 + e^2)^{\frac{1}{2}} = (.814)^{\frac{1}{2}} = .902$, the value used throughout the above analysis.

The estimate for the covariance (.186) is identical with that of Jencks, despite differences in the statistics. It is also much higher than the 14%–15% obtained directly from the Burks, Freeman, and Leahy studies. It should however be remembered that the environmental measures employed by these studies can only be approximate; it is possible that a more precise measure would yield a higher estimate for the covariance.

The Final Solution

As indicated earlier, four equations lead to the final solution. These are $h^2 + e^2 + 2hes = 1$ (Equation 4); $1.57gh(h + es) = .405$ (Equation 36); and $2hes = .186$ (Equation 38). The fourth equation is 32, which on substituting .692 for $r(1, 2)$ (Equation 19) and .57 for $r(\mathrm{F, M})$ becomes

$$.692 = h^2[.25 + .50g + 1.14g^2(h + es)^2] + 2hes + e^2 - \omega.$$

These four equations involve five unknowns: g, h, e, s, and ω. Since ω has a minimum value of .03 (the environmental variance within an identical twin pair) and, as Jencks et al.

Table 3
Values of h^2, e^2, *and Other Parameters Corresponding to Range of Values of* ω

	ω					
Parameter	.03	.04	.05	.06	.07	.08
$2hes$	.186	.186	.186	.186	.186	.186
g	.361	.368	.376	.383	.392	.400
h^2	.621	.607	.593	.579	.566	.552
e^2	.193	.207	.221	.234	.248	.262
s	.269	.263	.257	.253	.249	.245
$r(G_1, G_2)$	.553	.559	.566	.573	.580	.587

Note. ω = environmental variance within a fraternal twin pair.

stated, is not likely to exceed this figure by any great amount, I solve the four equations for g, h, e, and s, taking in turn values of .03, .04, . . . , .08 for ω. The results are presented in Table 3.

There remains the question of what value to choose for ω. There are two approaches to this question. The first concerns the difference between the correlations for fraternal twins and ordinary siblings: about .12 (.69 vs. .57). Jencks et al. (1972, pp. 289–209) had difficulty in trying to explain this difference. In particular, they considered the effect of the imperfections of the Stanford-Binet standardization. But probably at most these account for only .03, that is, about a quarter of the difference in correlation. In the end, Jencks et al. accepted .12 as largely representing the difference in the environmental variance within a pair for siblings vis-à-vis twins.

However, the principle of the variability of G with age is relevant here; the correlation between the genotypic values of ordinary siblings is less than that for fraternal twins—because of the age difference. A tentative analysis suggests that as much as .05 can be attributed to the G variability factor. Together with the effect of the faulty standardization, this suggests that possibly no more than .05 can be attributed to the difference in environmental variance for siblings vis-à-vis fraternal twins.

One would imagine that the difference in the environmental variance within a pair for identical and fraternal twins would be much less than that for fraternal twins and ordinary siblings—again because of the age factor—say,

Table 4
Estimates of Between-Families and Within-Family Components of Environmental Variance e^2 *for Range of Values of* ω

Variance estimates	ω .03	.04	.05	.06	.07	.08
	Present study					
Within-family component (v_R)[a]	.08	.09	.10	.11	.12	.13
Environmental variance (e^2)	.193	.207	.221	.234	.248	.262
Between-families component (v_B)	.113	.117	.121	.124	.128	.132
	Newman, Freeman, & Holzinger (1937) and Burt (1966)[d]					
e^2	.165[e]					
v_B[b]	.085	.075	.065	.055	.045	.035
% reduction in v_B due to selective placement[c]	24.5	35.9	46.3	55.6	64.9	73.5

[a] $v_R = \omega + .05$ (assuming that difference between fraternal twins and siblings for environmental variance within family is .05).
[b] Obtained by subtracting estimate of v_R (present study) from estimate of e^2 (Newman, Freeman, & Holzinger, 1937; Burt, 1966).
[c] Obtained by calculating 100 $(1 - R)$ when R = ratio of estimate of v_B (Newman, Freeman, & Holzinger, 1937; Burt, 1966) to estimate of v_B (present study).
[d] Studies used separated identical twins.
[e] Average of Newman, Freeman, and Holzinger (1937) and Burt (1966).

.01 or .02. It appears therefore that ω can be .04 or .05, but is unlikely to be more. It also follows that the environmental variance within a family is no more than .09 or .10 (i.e., 9% or 10% of the total IQ variance).

A second approach is provided by the data from studies in which identical twins are brought up separately or apart (ITA). If the twins are separated at an early age and if there is no selective placement, then the variance within a pair is an estimate of the total environmental variance (both between families and within family). The Newman et al. (1937) study gives a figure of 36.3 (corrected for error) for the variance within a pair, that is, 16.2% of the total IQ variance ($15^2 = 225$). The figure for Burt's (1966) group test IQ data is only slightly more. The weakness of ITA studies is the likelihood of a considerable degree of selective placement being involved in the assignment of the twins to their foster homes. For this reason, there is a general tendency to dismiss these studies as useless for the purpose of heritability analyses. This, however, is a mistake. An important point, which is overlooked, is that selective placement affects only between-families environmental differences; it does not affect the factors that contribute to within-family environmental differences. Consequently, the 16.5% derived from the Newman et al. and Burt studies can be regarded as the sum of two quantities: (a) the full environmental variance within a family and (b) a fraction of the environmental variance between families, depending on the degree of selective placement present. Having noted this point, one can perform the calculations presented in Table 4.

The first half of Table 4 sets out estimates of the within-family and between-families environmental variances obtained for the present study by assuming values of .03 to .08 for ω. By subtracting the within-family variance (v_R in Table 4) from the $e^2 = .165$ of the Newman et al. and Burt studies, one also obtains the estimates of the between-families variance for the ITA studies (v_B in Table 4). The differences between the figures in rows 3 and 5 can be attributed to selective placement, and in the last line of the table, the reduction

Table 5
Variation of r(*A*Cx, *NM; NSP*) *and Other Derived Functions With Chronological Age* (x)

Function	x: 2.17	4.25	7.0	8.2	9.3	13.5	16
$r(ACx, NM; NSP)$	.112	.243	.359	.390	.410	.445	.449
Corrected[a]	.034	.293	.372			.445	
$hr(G_{ACx}, Q_{NM})$[b]	.101	.219	.324	.352	.370	.402	.405
$r(Gx, G16)$[c]	.250	.541	.800	.869	.913	.991	1.000
$r(NCx, NC16; E \text{ constant})$[d]	.544	.722	.878	.920	.948	.995	1.000

Note. AC = foster child; NM = natural mother; NSP = no selective placement.
[a] The figures in this row are the four Skodak and Skeels (1949) correlations for the 63 adopted children and their natural mothers, corrected for attenuation, selective placement, and sampling error. The figures above (in the first row) were obtained by fitting a smooth curve to the four correlations (see Figure 1).
[b] Obtained by multiplying the figures in the first listed function by $\alpha = (h^2 + e^2)^{\frac{1}{2}}(=.902)$.
[c] Obtained by dividing the first listed function by .449 (see Equation 40).
[d] Obtained by multiplying the third listed function by $h^2(=.607)$ and adding .393 (see Equation 39).

in variance due to selective placement is given as a percentage. It seems very reasonable to assume that selective placement does not produce a reduction in variance greater than 50%. It follows, therefore, that ω should be in the range .03–.05.

In light of the two arguments presented above, I choose $\omega = .04$ in order to obtain the final solution. It follows from Tables 3 and 4 that the components of the IQ variance are 60.7% for heredity, 20.7% for environment (9% within a family and 11.7% between families), and 18.6% for covariance.

The solution is a compromise between Burt and Jencks. It supports Jencks et al.'s figure for covariance, but is closer to Burt (and Jensen) as far as the *relative* importance of heredity and environment is concerned (since the ratio 60.7:20.7 is approximately 75:25).

The components of the genetic variance. For $\omega = .04$, $r(G_1, G_2) = .559$ (see Table 3). It follows that the genetic variance between families is $.607 \times .599 = .339$, that is, 33.9% of the total IQ variance. The within-family component is $60.7 - 33.9 = 26.8\%$. However, in accordance with the principle of the variability of G with age, these results apply only to fraternal twins or siblings tested at the same age. For siblings of different ages, the between-families component will decrease with increase in age difference and the within-family component increase.

Turning now to Table 1, it can be shown that for $\omega = .04$, $p = (h + es)^2 r(\text{F, M}) = .460$ and $A = 2gp = 2(.368)(.460) = .330$. Also, $V_{A(\text{R})} = 2g(1 - A) = .487$ (48.7%), $V_{A(\text{AM})} = 2gA = .250$ (25%), and $V_D = 1 - 2g = .263$ (26.3%); that is, 26.3% of the genetic variance can be attributed to dominance.

A Check on the Final Solution

The method that has been described might be regarded as deriving from two hypotheses: (a) that the genotypic value G varies with age and (b) that this genetic variation with age can be measured through the variation of the correlation between the IQ of the adopted child and the IQ of the natural mother. Since the environmental component of the IQ variance is relatively small, a corollary to these hypotheses is that the variation of the correlation between IQ at a chronological age of 16 years and IQ at earlier ages can be explained mainly in terms of the variation of G with age, although changes in the individual's environment over the years must make a contribution.

Thus the solution, and therefore the rationale on which it is based, can be checked by using it to derive the correlations between the IQs of children at 16 years and their IQs at earlier ages (x) for the theoretical case in which there is no environmental change between the two ages. On the basis of the two hypotheses and the corollary just stated, I would expect the theoretical correlations to be slightly greater than the observed correlations, the difference

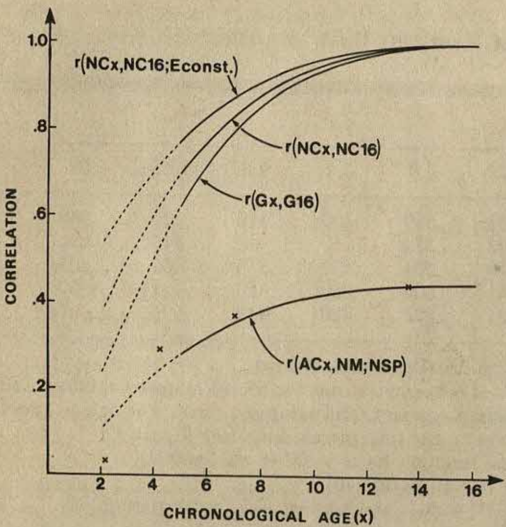

Figure 1. The variation of correlations with age (x); r(NCx, NC16; E constant) = correlation between IQ at age x and IQ at age 16 with no change in environment E; r(NCx, NC16) = correlation between IQ at age x and IQ at age 16 as obtained from empirical data; r(Gx, G16) = correlation between genotypic value at age x and genotypic value at age 16; r(ACx, NM; NSP) = correlation between IQ of adopted child at age x and IQ of natural mother with no selective placement.

(or drop) increasing with the size of the interval between the chronological ages of x and 16. Furthermore, I would expect the drop in correlation to remain well within the bounds of the total variance attributed to environmental differences (20.7%) if a reasonable amount of that variance is attributed to fairly constant environmental differences between families and also something to prenatal environmental differences. (Throughout the ensuing argument, it is understood that all correlations have been corrected for attenuation.)

For the theoretical case of no environmental change, one has for the child living with his own parents,

$$\begin{aligned} r(\mathrm{NC}x, \mathrm{NC16}; E \text{ constant}) &= E[(hG_{\mathrm{NC}x} + eE_{\mathrm{NC}x})(hG_{\mathrm{NC16}} + eE_{\mathrm{NC16}})] \\ &= h^2 r(G_{\mathrm{NC}x}, G_{\mathrm{NC16}}) + 2hes + e^2, \end{aligned}$$

making the usual assumption that covariance is constant for $x \geqslant 5$. Substituting the values for $2hes$ and e^2 provided by the final solution, one obtains

$$r(\mathrm{NC}x, \mathrm{NC16}; E \text{ constant}) = h^2 r(G_{\mathrm{NC}x}, G_{\mathrm{NC16}}) + .393. \quad (39)$$

It is necessary now to derive an expression for $r(G_{\mathrm{NC}x}, G_{\mathrm{NC16}})$ that will enable one to determine its value for any given x. I assume that

$$r(G_{\mathrm{NC}x}, Q_{\mathrm{NM}}) = r(G_{\mathrm{NC}x}, G_{\mathrm{NC16}})\, r(G_{\mathrm{NC16}}, Q_{\mathrm{NM}}),$$

whence

$$r(G_{\mathrm{NC}x}, G_{\mathrm{NC16}}) = \frac{r(G_{\mathrm{NC}x}, Q_{\mathrm{NM}})}{r(G_{\mathrm{NC16}}, Q_{\mathrm{NM}})} = \frac{r(\mathrm{AC}x, \mathrm{NM}; \mathrm{NSP})}{r(\mathrm{AC16}, \mathrm{NM}; \mathrm{NSP})} \quad (40)$$

(by Equation 29); that is, $r(G_{\mathrm{NC}x}, G_{\mathrm{NC16}}) = r(\mathrm{AC}x, \mathrm{NM}; \mathrm{NSP})/.449$, substituting the value for r(AC16, NM; NSP) obtained earlier.

By means of Equations 39 and 40, one can carry out the calculations summarized in Table 5. (The second function has been added to Table 5 to provide the values necessary for the calculation of selective placement corrections involved in the Burks and Leahy studies.)

Figure 1 gives a diagrammatic representation of the results. Also included in the diagram is the curve representing the variation of r(NCx, NC16) with x, as shown by empirical data (cf. Jensen, 1969, p. 18).

It is seen that the theoretical curve r(NCx, NC16; E constant) lies above the empirical curve r(NCx, NC16), the difference between the two correlations increasing as x decreases. However, the differences between the two correlations are quite small for $x \geqslant 5$. Thus at age $x = 5$, it amounts only to about .07.

It follows therefore that the correlation between IQ at earlier ages and IQ at age 16 can be explained mainly in terms of the variation of G with age and owes comparatively little to environmental changes with age. In other words, the analysis presented in this section provides a convincing check on the final solution and lends further support to the hypotheses and methods on which it was based.

References

Burks, B. S. The relative influence of nature and nurture upon mental development: A comparative

study of foster parent-foster child resemblance and true parent-true child resemblance. *27th Yearbook of the National Society for the Study of Education*, 1928, Pt. 1, 219–316.

Burt, C. The genetic determination of differences in intelligence: A study of monozygotic twins reared together and apart. *British Journal of Psychology*, 1966, *57*, 137–153.

Burt, C. *The gifted child*. London: Hodder & Stoughton, 1975.

Burt, C., & Howard, M. The multifactorial theory of inheritance and its application to intelligence. *British Journal of Statistical Psychology*, 1956, *9*, 95–131.

Cattell, R. B. The multiple abstract variance analysis equations and solutions: For nature–nurture research on continuous variables. *Psychological Review*, 1960, *67*, 353–372.

Falconer, D. S. *Introduction to quantitative genetics*. London: Oliver & Boyd, 1967.

Fisher, R. A. The correlations between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society* (*Edinburgh*), 1918, *52*, 399–433.

Freeman, F. N., Holzinger, K. J., & Mitchell, B. C. The influence of environment on the intelligence, school achievement and conduct of foster children. *27th Yearbook of the National Society for the Study of Education*, 1928, Pt. 1, 103–217.

Fulker, D. W. Applications of biometrical genetics to human behaviour. In Abeelen, J. H. F. van (Ed.), *The genetics of behaviour*. Amsterdam: North-Holland, 1974.

Gourlay, N. Heredity vs. environment: The effects of genetic variation with age. *British Journal of Educational Psychology*, 1978, *48*, 1–21.

Jencks, C. S., et al. *Inequality: A reassessment of the effect of family and schooling in America*. New York: Basic Books, 1972.

Jensen, A. R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, *39*, 1–123.

Jinks, J. L. & Eaves, L. J. IQ and inequality. *Nature*, 1974, *248*, 287–289.

Jinks, J. L., & Fulker, D. W. Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin*, 1970, *73*, 311–349.

Kamin, L. J. *The science and politics of I.Q.* New York: Wiley, 1974.

Layzer, D. Heritability analyses of IQ scores: Science or numerology? *Science*, 1974, *183*, 1259–1266.

Leahy, A. M. Nature-nurture and intelligence. *Genetic Psychology Monographs*, 1935, *17*, 236–308.

Mather, K. *Biometrical genetics: The study of continuous variations*. London: Methuen, 1949.

Newman, H. H., Freeman, F. N., & Holzinger, K. J. *Twins: A study of heredity and environment*. Chicago: University of Chicago Press, 1937.

Publications of the Scottish Council for Research in Education XXX. *The trend of Scottish intelligence*. London: University of London Press, 1949.

Rao, D. C., Morton, N. E., & Yee, S. Analysis of family resemblance: II. A linear model for familial correlation. *American Journal of Human Genetics*, 1974, *26*, 331–359.

Rao, D. C., Morton, N. E. & Yee, S. Resolution of cultural and biological inheritance by path analysis. *American Journal of Human Genetics*, 1976, *28*, 228–242.

Skodak, M., & Skeels, H. M. A final follow-up of 100 adopted children. *Journal of Genetic Psychology*, 1949, *75*, 85–125.

Wright, S. Statistical methods in biology. *Journal of the American Statistical Association*, 1931, *26*, 155–163. (Supplement)

(*Appendix follows*)

Appendix

Derivation of Basic Equations Used in the Analysis

1. For own children,

$$Q_{NCx} = hG_{NCx} + eE_{NCx} \quad \text{(Equation 20)},$$

where $h^2 + e^2 + 2hes = 1$ and $s = r_{GE}$, like h and e, is assumed to be constant for $x \geqslant 5$ years.

2. For foster children (no selective placement),

$$Q_{ACx} = (1/\alpha)(hG_{ACx} + eE_{ACx}) \quad \text{(Equation 21)},$$

where $\alpha = (h^2 + e^2)^{\frac{1}{2}}$. With no selective placement, there is no covariance. It is also assumed that the genetic and the environmental variances for foster children are the same as for own children.

In the case of selective placement, let

$$Q_{AC} = h'G_{ACx} + e'E_{ACx},$$

where $1 = h'^2 + e'^2 + e'^2 + 2h'e'r_{GE(\text{SP})}$.

Applying the same assumption as for no selective placement, it follows that

$$\frac{h'}{h} = \frac{e'}{e} = \frac{(h'^2 + e'^2)^{\frac{1}{2}}}{(h^2 + e^2)^{\frac{1}{2}}} = \frac{(1 - \text{cov}_{\text{AC}})^{\frac{1}{2}}}{\alpha},$$

where $\text{cov}_{\text{AC}} = 2h'e'r_{GE(\text{SP})}$.

In the case of the Burks, Leahy, and Skodak and Skeels studies, in which only a moderate degree of selective placement is involved, it can be shown that cov_{AC} lies in the range 0–.03. It follows that

$$h' \doteqdot h/\alpha \quad \text{and} \quad e' \doteqdot e/\alpha.$$

3. For foster child and foster parent with no selective placement (Equation 22),

$$Z = r(\text{AC, AP; NSP})$$
$$= E[(h'G_{ACx} + e'E_{ACx})Q_{AP}]$$
$$= E(e'E_{ACx}Q_{AP}),$$

since with no selective placement, there can be no correlation between Q_{AP} and G_{AC}; that is,

$$Z = (e/\alpha)r(E_{ACx}, Q_{AP}) = (e/\alpha)r(E_{NCx}, Q_{NP}),$$

assuming that parents treat foster children like their own children.

With the further assumption that the correlation of E_{NCx} with Q_{NP} is constant or approximately constant with age, it follows that

$$Z = (e/\alpha)r(E_{NC}, Q_{NP}).$$

4. For foster child and foster mother with selective placement (Equation 23),

$$r(\text{AC}x, \text{AM; SP})$$
$$= E[(h'G_{ACx} + e'E_{ACx})Q_{AM}]$$
$$= Z + (h/\alpha)r(G_{ACx}, Q_{AM}).$$

Assuming that selective placement takes place mainly through matching of mothers (natural and foster),

$$r(G_{ACx}, Q_{AM}) = r(G_{ACx}, Q_{NM})r(Q_{NM}, Q_{AM})$$
$$= r(G_{NCx}, Q_{NM})r(\text{NM, AM});$$

that is,

$$r(\text{AC}x, \text{AM; SP})$$
$$= Z + (h/\alpha)r(G_{NCx}, Q_{NM})r(\text{NM, AM}).$$

5. For parent and own child (Equations 25 and 26),

$$Y_x = r(\text{NC}_x, \text{NP})$$
$$= E[(hG_{NCx} + eE_{NCx})Q_{NP}]$$
$$= hr(G_{NCx}, Q_{NP}) + \alpha Z$$

(by Equation 22).

Assuming that at age $x = 16$, the G's of parents and children are comparable:

$$G_{NC16} = g(G_{NF} + G_{NM}) + \textit{random term};$$

that is,

$$r(G_{NC16}, Q_{NP}) = E[g(G_{NF} + G_{NM})Q_{NP}].$$

Making the same assumption as Jencks et al. (1972) and Fisher (1918), that assortative mating takes place through the phenotypic values (i.e., IQs),

$$E(G_{NF}Q_{NM})$$
$$= r(G_{NF}, Q_{NM})$$
$$= r(G_{NF}, Q_{NF})r(Q_{NF}, Q_{NM})$$
$$= (h + es)r(\text{F, M}) = E(G_{NM}, Q_{NF}),$$

whence

$$hr(G_{NC16}, Q_{NP}) = gh(h + es)[1 + r(\text{F, M})].$$

6. The derivation for covariance $2hes$ (Equation 27) is as follows: Implicit in Jencks et al.'s (1972, p. 268) path model is the assumption that the G of the child is correlated with the E entirely through the IQs of the parents. The same assumption is made here, except that in accordance with the principle of the variability of G with age, it is necessary that the G's of

parent and child be comparable, that is, belong to the same age level. It is assumed that this is so when the age of child (x) is 16. I therefore derive the covariance at age 16 in accordance with Jencks et al.'s assumption and then make the standard assumption that the value of the covariance at $x = 16$ is the same for all other ages at which $x \geqslant 5$.

$$G_{\mathrm{NC16}} = g(G_{\mathrm{NF}} + G_{\mathrm{NM}}) + \textit{random term}.$$

Also, applying simple regression analysis,

$$E_{\mathrm{NC16}} = k(Q_{\mathrm{NF}} + Q_{\mathrm{NM}}) + \textit{random term},$$

where

$$k = \frac{r(E_{\mathrm{NC}}, Q_{\mathrm{NP}})}{1 + r(\mathrm{F, M})} = \frac{\alpha Z}{e[1 + r(\mathrm{F, M})]}$$

(by Equation 22); that is,

$$\begin{aligned}2hes &= 2heE(G_{\mathrm{NC16}}, E_{\mathrm{NC16}})\\ &= 2heE[g(G_{\mathrm{NF}}+G_{\mathrm{NM}})k(Q_{\mathrm{NF}}+Q_{\mathrm{NM}})]\\ &= 2hegk2(h+es)[1+r(\mathrm{F, M})]\\ &= \frac{4\alpha Z}{1+r(\mathrm{F, M})}\{gh(h+es)[1+r(\mathrm{F, M})]\}\\ &= \frac{4\alpha Z}{1+r(\mathrm{F, M})}(Y_{16}-\alpha Z) \text{ by Explanation 5.}\end{aligned}$$

7. For foster child and natural mother (Equations 28 and 29),

$$\begin{aligned}&r(\mathrm{AC}x, \mathrm{NM}; \mathrm{SP})\\ &\quad = E[(h'G_{\mathrm{AC}x} + e'E_{\mathrm{AC}x})Q_{\mathrm{NM}}]\\ &\quad = (h/\alpha)r(G_{\mathrm{AC}x}, Q_{\mathrm{NM}}) + (e/\alpha)r(E_{\mathrm{AC}x}, Q_{\mathrm{NM}}).\end{aligned}$$

The first term on the right-hand side is $r(\mathrm{AC}x, \mathrm{NM}; \mathrm{NSP})$. From this, Equation 29 follows. Also,

$$\begin{aligned}r(E_{\mathrm{AC}x}, Q_{\mathrm{NM}}) &= r(E_{\mathrm{AC}x}, Q_{\mathrm{AM}})r(Q_{\mathrm{AM}}, Q_{\mathrm{NM}})\\ &= (\alpha Z/e)r(\mathrm{AM, NM})\end{aligned}$$

(Equation 22), making the usual assumption that selective placement takes place through matching of mothers. Hence,

$$(e/\alpha)r(E_{\mathrm{AC}x}, Q_{\mathrm{NM}}) = Zr(\mathrm{AM, NM}).$$

Received February 1, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 3, 616–626

Male and Female Spoken Language Differences: Stereotypes and Evidence

Adelaide Haas
Department of Speech Communication
State University of New York College at New Paltz

Male speech and female speech have been observed to differ in their form, topic, content, and use. Early writers were largely introspective in their analyses; more recent work has begun to provide empirical evidence. Men may be more loquacious and directive; they use more nonstandard forms, talk more about sports, money, and business, and more frequently refer to time, space, quantity, destructive action, perceptual attributes, physical movements, and objects. Women are often more supportive, polite, and expressive, talk more about home and family, and use more words implying feeling, evaluation, interpretation, and psychological state. A comprehensive theory of "genderlect" must include information about linguistic features under a multiplicity of conditions.

Both casual and serious observers of the human condition have long recognized that communication between the sexes is often frustrating. A possible cause of the difficulty is that men and women may in fact not really be speaking the same language (Jong, 1977; Reik, 1954).

Aspects of form, topic, content, and use[1] of spoken language have been identified as sex associated. Either men or women are more likely to produce specific utterances. Informal observations, speculations, and stereotypes in each category are discussed first. This presentation is followed by a report of empirical findings[2] from a variety of communication situations. Although reports of stereotypes and evidence for male and female spoken language differences do not always coincide, they both contribute to one's understanding of sex roles and communication.

This review is based on a dissertation submitted to Teachers College, Columbia University in partial fulfillment of the requirements for the PhD degree. Deep appreciation is expressed to Edward Mysak, Lois Bloom, and Mary Parlee for their useful suggestions, criticism, and encouragement.

Requests for reprints should be sent to Adelaide Haas, Department of Speech Communication, State University of New York, New Paltz, New York 12562.

Form

> The form of utterances can be described in terms of their acoustic, phonetic shape . . . in terms of the units of sound, or *phonology*, the units of meaning that are words or inflections, or *morphology*, and the ways in which units of meaning are combined with one another, or *syntax*. (Bloom & Lahey, 1978, p. 15)

Perhaps the most widespread belief about men's speech as compared with women's is that it is coarser and more direct. An early observer of style in language, Jesperson (1922/1949), observed women's speech to be generally more conservative than men's in the following ways: Men are readier to coin and use new terms, pun, utter slang expressions, and employ profanity and obscenity. Women, on the other hand,

> are shy of mentioning certain parts of the human body and certain natural functions by the direct and often rude denominations which men and especially young men prefer when among themselves. Women will therefore invent innocent and euphe-

[1] The categories *form, topic, content,* and *use* were suggested by Lois Bloom of Teachers College, Columbia University and are described in Bloom and Lahey (1978).

[2] Mary Parlee of Barnard College, Columbia University suggested an evaluative review, separating stereotypes from empirical findings.

mistic words and paraphrases which sometimes may in the long run come to be looked upon as the plain or blunt names and therefore in their turn have to be avoided and replaced by more decent words. (p. 245)

Reik (1954) affirmed that "we all know that there is a 'man talk' and a 'woman talk'" (p. 14). He observed that "men . . . will not hesitate to say 'Hell' or 'Damned.' . . . Women will rarely say 'It stinks' preferring to state that it has a bad smell" (p. 14).

More recently, Kramer (1974b) quoted the following: "*The New Seventeen* on people who use 'those four letter words': Boys find it especially repugnant when girls use those words. One boy described girls who use profanity as having nothing better to say" (p. 22).

Lakoff (1973) observed that men use stronger expletives such as *shit* and *damn*, whereas women use weaker or softer profanity such as *oh dear*, *goodness*, or *fudge*. Farb (1974) suggested that *dear me* and *gracious* are part of the female lexicon, and Ritti (1973) stated that most teachers of the sixth grade are well aware that young girls use far more "expressives" such as *oh* and *wow* than do the boys in their classes.

Farb wrote, "Nowadays young women use words that were formerly taboo for them with as much freedom as young men use them" (p. 50), but young men are not permitted the more euphemistic expressions. However, research on people's perceptions of language as either male or female suggests that the earlier stereotypes of coarse, free male language contrasted with euphemistic female forms still hold. Garcia-Zamor (Note 1) asked four boys and four girls in an upper-middle-class nursery school to indicate whether certain utterances were produced by a male or female doll; *shit* was seen by both boys and girls as male, and *drat* was seen by both as female. In a study of adults' stereotypes, Kramer (1974a) asked college students to determine whether various captions taken from *New Yorker* cartoons were uttered by males or females. Men in the cartoons were found to swear more than women and for more trivial reasons.

A careful review of the literature revealed no empirical studies of the comparative use of expletives. Profanity and obscenity do not readily submit to laboratory study. Documentation of this stereotype would require recording speech of female-only, male-only, and mixed-sex groups in various settings. The speakers should certainly not know they are being observed.

Reports by individual investigators writing about their own experiences (Key, 1975; Lakoff, 1975) strongly suggest that the form of expressives is sex associated. A possible explanation is that expressives "serve different functions for men and women. Males use them when they are angry or exasperated. . . . But women's exclamations are likely to convey enthusiasm" (Kramer, 1974a, p. 83).

The form of women's language is reputed to be more polite than the form of men's. Lakoff (1975) noted that "women are supposed to be particularly careful to say 'please' and 'thank you' . . . a woman who fails at these tasks is apt to be in more trouble than a man who does so" (p. 55). She speculated that "the more one compounds a request, the more characteristic it is of women's speech" (p. 19). An example of a doubly compound request is "Won't you please close the door?" (p. 18).

Only one very limited empirical study of politeness forms was found: 16 women born in Maine around 1900 used more politeness forms than 12 male counterparts when interviewed by college students (Hartman, 1976).

According to Austin (1965), high, oral sounds and giggling sounds are appropriate for females in courtship, whereas males produce low and nasal sounds. Coser (1960) recorded verbal interactions involving humor at 20 staff meetings of a mental hospital. She found that senior staff members (psychiatrists) made more jokes than junior staff members (paramedics) and that men made more witticisms than women (99 out of 103), but women often laughed harder. Coser suggested that this concurs with the sex roles of male authority and female receptivity. Haas (1978) similarly found that girls laughed more than boys in mixed-sex dyads.

Women are permitted to cry, as reflected in Key's (1975) observation that "if a female talks or cries into a pillow it's 'muffled sobbing'; if a male does the same, it's 'blubbering,' with negative connotations" (p. 109). Crying has been observed more frequently in girls than in boys. In an analysis of 200 quarrels of preschool children, Dawe (1934) found that 35.8% of the girls cried compared with only 20.2% of the boys.

Several writers (Labov, 1966; Levine & Crockett, 1966; Trudgill, 1972) have speculated that men use more slang expressions than women or even that slang is man's domain. Conklin (Note 2), however, observed that women's vernacular has not been studied and suggested a need to especially examine the dialect of all-female groups. Empirical phonological studies of -in versus -ing endings (Fischer, 1958), of -uh versus -er endings (Levine & Crockett, 1966; Wolfram, 1969), and of f, t, and th usage (Wolfram, 1969), show black females more likely to use standard forms than black men. Similar results were found in studies of pronominal apposition, as in "my brother he went to the park," and multiple negation (Shuy, Wolfram, & Riley, 1968). Garvey and Dickstein (1972) noted more nonstandard forms in the speech of six dyads of boys from four population groups (black, white, and low and middle socioeconomic status) than in matched girls.

Joffe (1948) noted sex differences in the vernacular of menstruation, including the greater use of color references by men and of personification by women. For example, men might say "she's waving the red flag," whereas women might refer to "having my friend." This finding was part of a larger study in New York City on attitudes and beliefs about menstruation.

Jesperson (1922/1949) believed women leave sentences unfinished or dangling more often than men. In an informal survey of television panel discussions, Bernard (1972) noted that women are more frequently interrupted than men. This may help explain the unfinished sentences. No empirical evidence for sex differences in sentence completeness has been noted. Zimmerman and West (1975), however, reported in a study of 11 male–female dyads that "virtually all the interruptions and overlaps are by the male speakers (98% and 100% respectively)" (p. 115). They further noted that not one of the women who were interrupted protested. Similar results were reported by Eakins and Eakins (1976).

Women have long had the reputation for being more loquacious than men: *"Où femme il y a, silence il n'y a"* (Where there's woman, there's no silence.) "The tongue is the sword of a woman, and she never lets it become rusty" (China). "The North Sea will sooner be found wanting in water than a woman at a loss for a word" (Jutland; cited in Jesperson, 1922/1949, p. 253). Jesperson believed that

> the superior readiness of speech of women is a concomitant of the fact that their vocabulary is smaller than that of men. But this again is connected with another indubitable fact, that women do not reach the same extreme points as men, but are nearer the average in most respects. (p. 253)

He gave many examples of how women are supposed to talk ahead of thinking, to talk more than men.

Lakoff (1975) informally observed longer sentence forms in women than in men, possibly resulting in the impression of more speech. For example, women are more likely to compound a request: "Will you help me with these groceries, please?" is more characteristic of women than "Help me" or even "Please help me with these groceries." Empirical evidence, however, suggests that at least under certain conditions women's sentences are shorter than men's. For example, at professional conferences, the mean time used by women asking a question was reported to be less than half that used by men (Swacker, 1975).

Studies of sex differences in length of utterance in children indicate that girls are significantly superior to boys at various matched age levels in mean length of utterance (Winitz, 1959). Maccoby (1966, p. 335) reported similar results in her summary of 19 studies. Garvey and Ben Debba (1974), however, found no sex differences in words per utterance among same-sex or mixed-sex dyads ranging in age from $3\frac{1}{2}$ to $5\frac{3}{4}$ years

and participating in free-play testing situations. In considering mean length of utterance of children, language maturation must be considered a factor, since utterances normally become longer as skill in language increases and most studies show that girls develop language facility earlier than boys. Limited evidence, then, suggests that although in early childhood female sentences are longer than those of males, by adulthood the reverse may be true.

Mixed results have been reported in studies of verbosity (Maccoby, 1966, p. 335). In a task involving adults' responses to picture stimuli, Wood (1966) concluded that men tend to use more words than women in responding to a given stimulus. Like results in similar situations were found by Argyle, Lalljee, and Cook (1968) and Swacker (1975). Cherry's (Note 3) review of 11 studies dealing with children's quantity of speech reported that girls tended to exceed boys in this dimension in 6 of the studies. No differences were noted in 4 studies.

The participants in a communication influence quantity of verbalization. In mixed-sex groups, men tend to talk more than women (Argyle et al., 1968; Bernard, 1972).

Among children the composition of the communication group also seems to affect verbosity. Mueller (1972), in a study of "the maintenance of verbal exchanges between young children" (ages 3½–5 years) found that "boys talked significantly more than girls" (p. 933) in a free-play situation to same-sex peers. Brownell and Smith (1973), however, reported more verbal productivity among 4-year-old girls in comparison with same-age boys in mixed-sex dyads, triads, and small groups. In preschool children, then, boys have been found to talk more to boys and girls to talk more in mixed-sex groups—the reverse of the adult pattern.

Entwisle and Garvey (1972) reported sex differences in verbal productivity among Baltimore children, with girls more productive than boys; note that this finding is most marked among those of lower social class. Possibly no real difference exists in the quantity of talk that is produced by men and women, but "girls are not *supposed* to talk as much as a man" (Kramer, 1974b, p. 17).

In sum, the stereotype clearly shows women to be more verbose than men. Empirical evidence is mixed. Girls seem to talk somewhat more than boys, but adult women, especially in the company of men, have been found to talk less than their male companions.

Topic

Topic refers to the subject matter of the spoken utterance, to what the conversation is about.

Kramer (1974b) captured much of the folklore related to topics of male and female conversations through her study based on *New Yorker* cartoons:

> Men hold forth with authority on business, politics, legal matters, taxes, age, household expenses, electronic bugging, church collections, kissing, baseball, human relations, health and—women's speech. Women discuss social life, books, food and drink, pornography, life's troubles, caring for a husband, social work, age, and life-style. Several of the students who rated the cartoon captions said they considered all statements about economics, business or jobs to be male. (p. 83)

The interviews by Komarovsky (1967) suggest similar stereotypes in blue-collar families. One 28-year-old wife commented that "[men] think we [women] are silly and talk too much. They think that women gossip a lot and they are against it" (p. 150). A 36-year-old husband noted that women want "to talk about kidstuff and trivia like Mrs. X had her tooth pulled out" (p. 150). Women reported that they enjoyed talking about the family and social problems. Both sexes acknowledged that men prefer to talk about cars, sports, work, motorcycles, and local politics.

Klein's (1971) observations of the working class in England are similar:

> Just as men in the clubs talk mainly about their work and secondly about sport and *never* about their homes and families, so do their wives talk first of all about *their* work, i.e.: their homes and families, and secondly within the range of things with which they are all immediately familiar. (p. 73)

In mixed-sex conversations the impression is that women initiate topics that are rarely followed through by men (Bernard, 1972; Chesler, 1972).

Three studies in the 1920s of conversational topics using tape-recorded fragments of conversations on city streets are of interest. Moore (1922) recorded 174 conversations in New York City and reported that man-to-man topics included money and business (48%), amusements or sports (14%), and other men (13%). Woman-to-woman topics were men (44%), clothing or decoration (23%), and other women (16%). Male-to-female topics were amusements or sports (25%) and money and business (22%). Women talked to men about other men (22%) and other women (13%).

M. H. Landis and Burtt (1924) conducted a similar study in Columbus, Ohio and recorded 481 conversations. Their findings concur with Moore's. Men talked to men about business and money 49% of the time, sports or amusements 15% of the time, and other men 13% of the time. Women talked to women about men (22%), clothing or decoration (19%), and other women (15%). Women talked about people in 37% of the conversations. Man-to-woman topics included amusement and sports (25%), money and business (19%), and themselves (23%). Women talked to men about amusements or sports (24%), clothing or decoration (17%), and themselves (17%).

In 1927, C. Landis analyzed 200 London conversations. The all-male topics were similar to those in New York City and Columbus, but the women talked about a wider variety of topics among themselves. Landis suggested that in mixed-sex conversations, "the Englishman when talking to a feminine companion adapts his conversation to her interests while American women adapt their conversations to the interests of their masculine companions" (p. 357).

In a study of "the women of the telephone company," Langer (1970a, 1970b) reported that men discussed politics among themselves, whereas women avoided religion and politics in their conversations.

Mulcahy (1973), using a self-disclosure questionnaire with 97 adolescents, reported that female same-sex disclosure was greater than male same-sex disclosure. Major topics for girls were "tastes, interests, and personality" (p. 343); for boys high disclosure clustered about "tastes, and interests, work (studies), and attitudes and opinions" (p. 343). "The lowest disclosure area for males was Body, whereas it was Money for females" (p. 354).

Sause (1976) reported that kindergarten girls made more reference to the female role than did kindergarten boys, and this was the only category that girls referred to more than boys in this study of 144 subjects. Boys talked more about family and home environment, recreation, other people, and animals, but the differences were not significant. Utterances were all to a male examiner who encouraged the children to talk about two stimulus objects—an irregularly shaped block and a toy fire engine.

Knowledge of conversational topics is limited. Although the evidence supports the stereotype that women talk more about people and men more about money, business, and politics, the studies date back to the 1920s. Times have been changing!

Content

Content refers to the "categorization of the topics that are encoded in messages," such as "object in general," "actions in general," and the "possession relation in general" (Bloom & Lahey, 1978, p. 11). Content differs from topic, since topic refers to particular objects, events, and ideas, whereas content refers to the more general concept of how the topic is referenced.

Women's language is more emotional and evaluative than men's according to the stereotype (Jesperson, 1922/1949; Kramer, 1974a; Lakoff, 1975; Pei, 1969; Reik, 1954). Jesperson wrote of women's fondness for hyperbole and their greater use of adverbs of intensity such as *awful, pretty, terribly nice, quite,* and *so.* These all suggest value judgments. Reik believed terms such as *darling, divine, sweet, adorable, I could just scream,*

I nearly fainted, and *I died laughing* are female associated. Pei observed "extravagant adjectives" such as *wonderful, heavenly, divine*, and *dreamy* in women's speech. Again the focus is on emotional value judgment.

Lakoff's (1975) list of female adjectives includes *adorable, charming, lovely*, and *divine*. Male adjectives are *great, terrific* and *neat*. Kramer (1974a) suggested that "words of approval" (p. 22) such as *nice, pretty, darling, charming, sweet, lovely, cute*, and *precious* are used more frequently by women.

Hartman (1976) tested and supported Lakoff's hypothesis that women use evaluative adjectives more than men. In her study of 70-year-old native Maine men and women, she found that women compared with men used many more words such as *lovely, delightful, wonderful, nice, pretty, pathetic, pretty little, smartly uniformed, cute, dearest, gentle, gaily, beautifully, lovelies, very very, devoted, meek, perfectly wonderful*, and *stylish*. Most women used *awful* and *pretty* to mean *very* and *so*.

Wood (1966) analyzed the speech of 36 college students (18 men and 18 women) as they described photographs of a man's face. She found that males referred more directly to what was actually in the picture. Females were more interpretative and tended to be more subjective in their descriptions. Barron's (1971) study of speech by teachers and pupils during regular classroom activities showed patterns similar to those reported by Wood. Through an analysis of the grammatical cases of speaker's utterances, Barron found that women used more participative and purposive cases and men used more instrumental and objective cases. Specifically, women talked more about how people felt and why they behaved in certain ways. Men's speech focused more on objects and actions related to these objects.

Gleser, Gottschalk, and Watkins (1959) studied the speech of 90 white adult men and women who were asked by a male examiner to tell about "any interesting or dramatic life experiences you have had" (p. 183). As did the other studies, this investigation revealed that women used significantly more words implying feeling, emotion, or motivation (whether positive, negative, or neutral); they made more self-references and used more auxiliary words and negations. Male subjects referred more to time, space, quantity, and destructive action. This can be viewed as supporting Eble's (Note 4) suggestion that terms of hostility are more associated with men.

Physical movement was more frequently referenced by kindergarten boys than by girls (Sause, 1976). Boys also used significantly more words classified as self, space, quantity, good, bad, and negative words. Garcia and Frosch (1976) also found that males talked more about spatial relations than females. Their subjects were 40 black, Anglo, and Spanish-speaking adults, ranging in age from 18 to 65 years, who were asked to respond to two pictures (one "female room" and one "male outdoors scene") from current magazines. Females described items in terms of patterns and colors more than did males. Also of interest was the observation that "each [sex] group went into immediate detail when describing the visual which was stereotyped to their sex group, but paused to 'orient themselves' to the environment when approaching the other visual" (p. 68).

Comparative use of adjectives was studied by Kramer (1974b), Brandis and Henderson (1970), and Entwisle and Garvey (1972). College students writing descriptions of black and white photographs did not differ in the type or number of prenominal adjectives used or in the number or variety of -ly adverbs (Kramer, 1974b). However, according to the studies by Brandis and Henderson and Entwisle and Garvey, girls use more adjectives than boys. The Brandis and Henderson study was on spoken language by 5-year-old working-class British children; the Entwisle and Garvey study was based on the written language of ninth graders when asked to write imaginative stories after viewing four stimulus pictures.

Garvey and Dickstein (1972) found that fifth-grade black and white boys of low and middle socioeconomic status used the possessive construction more frequently than females of the same age, race, and socioeco-

nomic status during oral communication involving problem-solving tasks.

The stereotype of the content of spoken language, then, points to positive value judgments as female marked and hostile judgments as male marked. The empirical evidence suggests that the content of adult female speech includes more words implying feeling, auxiliary words, negations, evaluative adjectives, interpretations, psychological state verbs, and purposive cases. Adult males use more terms referring to time, space, quantity, destructive action, and perceptual attributes and more objective cases. Boys have been reported to use more words related to self, space, quantity, good, bad, negation, and possession. It is likely that girls use more adjectives. Studies of adult use of adjectives show mixed results.

Use

"Language use consists of the socially and cognitively determined selection of behaviors according to the goals of the speaker and the context of the situation" (Bloom & Lahey, 1978, p. 20).

Bernard (1972) suggested that "instrumental" talk is male associated. Men are stereotyped as the conveyors of information and fact. Women "tend to be handicapped in fact-anchored talk. . . . They are . . . less likely to have a hard, factual background, less in contact with the world of knowledge" (p. 153). The male instrumental style includes lecturing, argument, and debate. This has not been empirically documented to date.

Assertiveness was observed as part of the male stereotype by Kramer (1974b) in her study of cartoon captions. Lakoff (1975) suggested that women's speech is nonassertive. This concept has been developed by other writers. Kuykendall (Note 5) wrote that "clean, effective vigorous speech and writing is just what women, *qua* women, learn not to produce so as not to appear too assertive and so to offend" (p. 4). Furthermore, "Assertion of competence and power by a female is regarded as deviant behavior so that she becomes the recipient of social sanctions" (Unger, Note 6, p. 43). Wolman and Frank (1975) observed that in a professional peer group a woman was labeled *bitchy* or *manipulative* when her behavior was assertive and directive. Nursery school children also believe that competitive and aggressive language is appropriate for males only, as demonstrated by a study in which boys and girls were asked to ascribe various uttered sentences to a girl or boy doll (Garcia-Zamor, Note 1). Dawe (1934) found that when nursery school children quarreled, boys were assertive by threatening and forbidding more often than girls.

Tentativeness has been stereotyped as female. Lakoff (1975) suggested that tag questions (e.g., "It's cold, isn't it?") are used far more often by women than by men. This form of question avoids assertion and gives the addressee the option of agreeing or disagreeing. Women's speech is said to be "hedge marked."

Empirical evidence is mixed. Hartman (1976) reported that tentativeness was clearly female associated among the 70-year-old Maine natives whose speech she studied. This was revealed in the women's greater production of qualifiers such as *perhaps, I suppose, I just feel, probably,* and *as I interpret it* and tag questions such as "Well, most people would say marriage, wouldn't they?" and "It was grandmother, wasn't it?" Swacker (1975) found that female college students indicated approximation when using numbers ("about six books"), whereas only one male used the tentative form in a task requiring the description of three pictures by Albrecht Dürer. However, in dyadic conversations of college students, Hirschman (Note 7) found no difference between the sexes in the overall proportion of qualifiers such as *maybe, probably, I think,* and *I guess.* In a somewhat larger study, Hirschman (Note 8) found that males uttered *I think* twice as much as females. (*I think* is usually considered a qualifier, but Hirschman suggested that it served primarily as a way for more assertive speakers to present their opinions.) Loban (Note 9) reported that expressions of tentativeness including supposition, hypothesis, and conditionality are associated with effective users of language from

kindergarten through sixth grade. Hass and Wepman (1973) similarly found that uncertainty increased as a function of age in children 5 to 13 years old and noted that "there are many fine points about the uncertainty scores [with regard to the Age × Sex interaction] that demand further investigation" (p. 305). Baumann (1976) analyzed 7½ hours of tape of adults in various settings for confirmatory tag questions and qualifying prefatory statements. She found only 20 examples altogether and no sex-associated use.

Men and women may make requests in different ways. Lakoff (1975) observed that women state requests and men issue commands. Hennessee and Nicholson (1972) reported that in over 1,000 television commercials, men gave almost 90% of the directives, that is, the advice or commands to buy a particular product. In a naturalistic study of the conversations of a single married couple, Soskin and John (1963) reported that the husband gave far more directives than the wife. In one critical situation when they were rowing and the boat capsized, mainly the husband gave regulative statements such as demands, suggestions, and prohibitions.

Hirschman (Note 8) tested the hypothesis that women are more supportive than men. No overall differences were found between the college men and college women studied, although females used 'mm hmm" significantly more than males and most of these utterances occurred in female-to-female conversations. In mock jury deliberations, Strodtbeck and Mann (1956) reported that women agreed, concurred, complied, accepted, and supported other speakers almost twice as much as men did. Similarly, women were antagonistic or offensive half as often as men. Conversely, men were more assertive. Supportive behavior can be inferred from the emotional sensitivity Alvy (1973) reported to be more characteristic of grade-school girls than of boys of lower, middle and upper socioeconomic status in an experiment of listener-adapted communication.

In use, then, men's speech reputedly serves to lecture, argue, debate, assert, and command. Women's speech is stereotyped as nonassertive, tentative, and supportive. Limited evidence confirms that males are more assertive and issue more directives; females are often more tentative and supportive.

Conclusions and Implications

Do male and female spoken language differences exist? The stereotypes abound, and evidence has been accumulating, especially since the beginning of this decade.

Women's speech is said to contain more euphemisms, politeness forms, apology, laughter, crying, and unfinished sentences. They are reputed to talk more about home and family and to be more emotional and positively evaluative. Further, women's speech is stereotyped as nonassertive, tentative, and supportive. Women are also said to talk more than men.

Men, on the other hand, are reputed to use more slang, profanity, and obscenity and to talk more about sports, money, and business. They are reputed to make more hostile judgments and to use language to lecture, argue, debate, assert, and command.

Empirical evidence is less clear, partly because studies can only sample limited populations in specific situations. Further, sex differences in American English are only statistical differences. No feature of spoken American English is used exclusively by one sex or the other. In general, however, empirical studies of form confirm that males use more nonstandard forms than females and that females laugh and cry more. Older Maine women, at least, are more polite, and sixth-grade girls claim they use more expressives. Contrary to the stereotype, adult men have been found to be more loquacious, but it is unclear whether boys or girls are more verbose. Studies from the 1920s support the stereotype that men talk more about money, business, and politics and that women talk more about home and family. The empirical evidence supports the stereotype of content differences in men's and women's speech. Various studies found that women use more emotional language and men focus more on perceptual attributes and destructive action. The males studied were generally

more assertive and directive than the women. One study found that women are more supportive than men, and the results of research on tentativeness are mixed.

Are these isolated, unrelated variations in speech, or is there a logical clustering that points to "systems of co-occurring, sex-linked signals," or "genderlects," as Kramer (1974b, p. 14) proposed?

If, in fact, one can say that there is a male speech style and a female speech style, then rules and restrictions can be written for each much in the way that grammatical structures are described. This task is complicated by two major observations: (a) Sex differences in spoken language that have been identified in English are sex preferential as opposed to sex exclusive (Bodine, 1975); that is, there is no evidence that any linguistic feature is used exclusively by one sex in our society; variations have been found only in frequency of production. (b) Sex is not the only variable to influence speech style. There is a complex interaction of personal characteristics such as sex, age, education, occupation, geographical region, ethnic background, and socioeconomic status and contextual factors such as communication, situation, environment, and participants.

Despite these complications, a start has been made at constructing a grammar of style for men's and women's language (Lakoff, Note 10). Lakoff focused on women's style and suggested that it is basically one of deference. She suggested that the various phonological and lexical forms and the syntactic–pragmatic features identified as occurring more often in women's speech add up to a pattern of deference. However, deference alone does not make a woman's style. Other characteristics of the individual and the context combine to form the complete style. Lakoff pointed to a need to learn which styles can coexist and which cannot. Even more important is the need to know which sex-associated spoken language features are real and to document conditions under which they occur.

Communication can be viewed as a microcosm of social behavior. Much of human interaction occurs at the linguistic level. As Gumperz and Hymes (1972) pointed out,

> If sociolinguistic research often begins as an extension of linguistics, it must end as an intension of the social sciences—but in the idiom of disciplines that is only to say that it changes from a way of studying language to a way of studying man as a social being. (p. 466)

The stereotypes and evidence discussed in this article have significant implications for the power structure between the sexes and indeed the psyche of both men and women. Future researchers need to be sensitive to situations in which they observe sex-associated speech and to be cautious of making premature judgments. In any event, there is little doubt that recent interest in gender and language will continue to generate worthwhile exploration into this topic. Clinicians and theoreticians alike will thereby increase their understanding of this important dimension of human communication.

Reference Notes

1. Garcia-Zamor, M. A. *Child awareness of sex-role distinctions in language use.* Paper presented at the meeting of the Linguistic Society of America, San Diego, Calif., December 1973.
2. Conklin, N. F. *Perspectives on the dialects of women.* Paper presented at the meeting of the American Dialect Society, 1973.
3. Cherry, L. J. *Sex differences in child speech: McCarthy revisited.* Princeon, N.J.: Educational Testing Service, February 1975.
4. Eble, C. C. *How the speech of some is more equal than others.* Paper presented at the meeting of the Southeastern Conference on Linguistics, University of North Carolina at Chapel Hill, 1972.
5. Kuykendall, E. *Sexism in language.* Unpublished manuscript, State University of New York College at New Paltz, Department of Philosophy, 1976.
6. Unger, R. K. *Status, power and gender: An examination of parallelisms.* Paper presented at the Conference on New Directions for Research on Women, Madison, Wis., May–June 1975.
7. Hirschman, L. *Female-male differences in conversational interaction.* Paper presented at the meeting of the Linguistic Society of America, San Diego, Calif., December 1973.
8. Hirschman, L. *Analysis of supportive and assertive behavior in conversations.* Paper presented at the meeting of the Linguistic Society of America, July 1974.

9. Loban, W. D. *The language of elementary school children* (Report No. 1). Champaign, Ill.: National Council of Teachers of English, 1963.
10. Lakoff, R. *Women's styles of speaking: Their psychological significance.* Paper presented at the Conference on Women's Language, Graduate School and University Center of the City University of New York, April 1977.

References

Alvy, K. T. The development of listener adapted communications in grade-school children from different social-class backgrounds. *Genetic Psychology Monographs,* 1973, *87,* 33–104.

Argyle, M., Lalljee, M., & Cook, M. The effects of visibility on interaction in a dyad. *Human Relations,* 1968, *21,* 3–17.

Austin, W. M. Some social aspects of paralanguage. *Canadian Journal of Linguistics,* 1965, *11,* 31–39.

Barron, N. Sex-typed language: The production of grammatical cases. *Acta Sociologica,* 1971, *14,* 24–72.

Baumann, M. Two features of "women's speech?" In B. L. Dubois & I. Crouch (Eds.), *The sociology of the languages of American women.* San Antonio, Tex.: Trinity University Press, 1976.

Bernard, J. *The sex game.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Bloom, L., & Lahey, M. *Language development and language disorders.* New York: Wiley, 1978.

Bodine, A. Sex differentiation in language. In B. Thorne & N. Henley (Eds.), *Language and sex: Difference and dominance.* Rowley, Mass.: Newbury House, 1975.

Brandis, W., & Henderson, D. *Social class, language and communication.* London: Routledge & Kegan Paul, 1970.

Brownell, W., & Smith, D. R. Communication patterns, sex and length of verbalizations in the speech of four-year old children. *Speech Monographs,* 1973, *40,* 310–316.

Chesler, P. *Women and madness.* Garden City, N.Y.: Doubleday, 1972.

Coser, R. L. Laughter among colleagues. *Psychiatry,* 1960, *23,* 81–95.

Dawe, H. C. An analysis of 200 quarrels of preschool children. *Child Development,* 1934, *5,* 139–157.

Eakins, B., & Eakins, G. Verbal turn-taking and exchanges in faculty dialogue. In B. L. Dubois & I. Crouch (Eds.), *The sociology of the languages of American women.* San Antonio, Tex.: Trinity University Press, 1976.

Entwisle, D. R., & Garvey, C. Verbal productivity and adjective usage. *Language and Speech,* 1972, *15,* 288–298.

Farb, P. *Word play: What happens when people talk.* New York: Knopf, 1974.

Fischer, J. L. Social influences on the choice of a linguistic variant. *Word,* 1958, *14,* 47–56.

Garcia, G. N., & Frosch, S. F. Sex, color and money: Who's perceiving what? Or men and women: Where did all the differences go (to?)? In B. L. Dubois & I. Crouch (Eds.), *The sociology of the languages of American women.* San Antonio, Tex.: Trinity University Press, 1976.

Garvey, C., & Ben Debba M. Effects of age, sex, and partner on children's dyadic speech. *Child Development,* 1974, *45,* 1159–1161.

Garvey, C., & Dickstein, E. Levels of analysis and social class differences in language. *Language and Speech,* 1972, *15,* 375–384.

Gleser, G. C., Gottschalk, L. A., & Watkins, J. The relationship of sex and intelligence to choice of words: A normative study of verbal behavior. *Journal of Clinical Psychology,* 1959, *15,* 182–191.

Gumperz, J. J., & Hymes, D. H. *Directions in sociolinguistics: The ethnography of communication.* New York: Holt, Rinehart & Winston, 1972.

Haas, A. *Production of sex-associated features of spoken language by four-, eight-, and twelve-year old boys and girls* (Doctoral dissertation, Columbia University, Teachers College, 1977). *Dissertation Abstracts International,* 1978, *39,* 23A.

Hartman, M. A descriptive study of the language of men and women born in Maine around 1900 as it reflects the Lakoff hypotheses in "Language and woman's place." In B. L. Dubois & I. Crouch (Eds.), *The sociology of the languages of American women.* San Antonio, Tex.: Trinity University Press, 1976.

Hass, W. A., & Wepman, J. M. Constructional variety in the spoken language of school children. *Journal of Genetic Psychology,* 1973, *122,* 297–308.

Hennessee, J., & Nicholson, J. NOW says: TV commercials insult women. *New York Times Magazine,* May 28, 1972, pp. 12–13; 48–51.

Jesperson, O. *Language.* New York: Macmillan, 1949. (Originally published, 1922.)

Joffe, N. F. The vernacular of menstruation. *Word,* 1948, *4,* 181–186.

Jong, E. *How to save your own life.* New York: Holt, Rinehart & Winston, 1977.

Key, M. R. *Male/female language.* Metuchen, N.J.: Scarecrow Press, 1975.

Klein, J. The family in "traditional" working-class England. In M. Anderson (Ed.), *Sociology of the family.* Baltimore, Md.: Penguin Books, 1971.

Komarovsky, M. *Blue-collar marriage.* New York: Vintage Books, 1967.

Kramer, C. Folklinguistics. *Psychology Today,* June 1974, pp. 82–85. (a)

Kramer, C. Women's speech: Separate but unequal? *Quarterly Journal of Speech,* 1974, *60,* 14–24. (b)

Labov, W. *The social stratification of English in New York City.* Washington, D.C.: Center for Applied Linguistics, 1966.

Lakoff, R. Language and woman's place. *Language in Society,* 1973, *2,* 45–80.

Lakoff, R. *Language and woman's place.* New York: Colophon/Harper & Row, 1975.

Landis, C. National differences in conversations. *Journal of Abnormal and Social Psychology,* 1927, *21,* 354–357.

Landis, M. H., & Burtt, H. E. A study of conversations. *Journal of Comparative Psychology,* 1924, *4,* 81–89.

Langer, E. The women of the telephone company: Part 1. *New York Review of Books,* March 12, 1970, pp. 16; 18; 20–24. (a)

Langer, E. The women of the telephone company: Part 2. *New York Review of Books,* March 26, 1970, pp. 14; 16–22. (b)

Levine, L., & Crockett, H. J. Speech variation in a Piedmont community: Post-vocalic r. In S. Lieberson (Ed.), *Explorations in sociolinguistics.* The Hague, Netherlands: Mouton, 1966.

Maccoby, E. E. (Ed.). *The development of sex differences.* Stanford, Calif.: Stanford University Press, 1966.

Moore, H. T. Further data concerning sex differences. *Journal of Abnormal and Social Psychology,* 1922, *4,* 81–89.

Mueller, E. The maintenance of verbal exchanges between young children. *Child Development,* 1972, *43,* 930–938.

Mulcahy, G. A. Sex differences in patterns of self-disclosure among adolescents: A developmental perspective. *Journal of Youth and Adolescence,* 1973, *4,* 343–356.

Pei, M. *Words in sheep's clothing.* New York: Hawthorn Books, 1969.

Reik, T. Men and women speak different languages. *Psychoanalysis,* 1954, *2,* 3–15.

Ritti, A. *Social functions of children's speech* (Doctoral dissertation, Columbia University, Teachers College, 1972. *Dissertation Abstracts International,* 1973, *34,* 2289B.

Sause, E. F. Computer content analysis of sex differences in the language of children. *Journal of Psycholinguistic Research,* 1976, *5,* 311–324.

Shuy, R. W., Wolfram, W. A., & Riley, W. K. *Field techniques in an urban language study.* Washington, D.C.: Center for Applied Linguistics, 1968.

Soskin, W. F., & John, V. P. The study of spontaneous talk. In R. Barker (Ed.), *The stream of behavior.* New York: Appleton-Century-Crofts, 1963.

Strodtbeck, F. L., & Mann, R. D. Sex role differentiation in jury deliberations. *Sociometry,* 1956, *19,* 3–11.

Swacker, M. The sex of the speaker as a sociolinguistic variable. In B. Thorne & N. Henley (Eds.), *Language and sex: Difference and dominance.* Rowley, Mass.: Newbury House, 1975.

Trudgill, P. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society,* 1972, *1,* 179–195.

Winitz, H. Language skills of male and female kindergarten children. *Journal of Speech and Hearing Research,* 1959, *2,* 377–391.

Wolfram, W. *A sociolinguistic description of Detroit Negro speech.* Washington, D.C.: Center for Applied Linguistics, 1969.

Wolman, C., & Frank, H. The solo woman in a professional peer group. *American Journal of Orthopsychiatry,* 1975, *45,* 164–171.

Wood, M. The influence of sex and knowledge of communication effectiveness on spontaneous speech. *Word,* 1966, *22,* 112–137.

Zimmerman, D. H., & West, C. Sex roles, interruptions and silences in conversation. In B. Thorne & N. Henley (Eds.), *Language and sex: Difference and dominance.* Rowley, Mass.: Newbury House, 1975.

Received February 3, 1978 ▪

Psychological Bulletin
1979, Vol. 86, No. 3, 627–637

Habituation Model of Systematic Desensitization

Fraser N. Watts
King's College Hospital
University of London, London, England

The relevance of habituation as a model for response decrement in desensitization is considered. A discussion of the relationship between habituation and extinction leads to the view that there are no sound reasons for explaining desensitization as an extinction rather than as a habituation phenomenon. The maximal habituation theory of desensitization proposed by Lader and Mathews is discussed and relevant evidence reviewed. Finally, a revised habituation theory of desensitization, based on the dual-process theory of habituation, is elaborated, and the role in desensitization of relaxation, stimulus intensity, stimulus lengths, and interstimulus interval lengths are discussed in the context of this theory. It is suggested that relaxation and an incremental stimulus hierarchy may reduce sensitization rather than facilitate habituation.

More than 10 years ago, Lader and Wing (1966) proposed a habituation theory of desensitization. The theory was subsequently elaborated by Lader and Mathews (1968) and has become known as the *maximal habituation* theory of desensitization. Though it has not won general acceptance, it has stimulated a significant body of research and has increased our knowledge about the processes at work in desensitization. It has thus passed at least one test of a useful theory. However, it is now clear that this version of the habituation theory of densensitization has a number of weaknesses. It is the purpose of this article to review relevant research and to elaborate a reformulation of the habituation theory of desensitization in the light of the dual-process habituation theory (Groves & Thompson, 1970) that has been developed since the maximal habituation theory of desensitization was proposed.

Desensitization: Habituation or Extinction?

Habituation can be defined as the waning of a response to a stimulus that occurs when the stimulus is repeatedly presented. It is usually said to be a central nervous system process, which distinguishes it from other decremental processes such as fatigue. It is also said to apply only to the decrement of unconditioned responses. This distinguishes it from extinction, which applies to conditioned responses. Though the habituation of a wide variety of responses has been described in the literature, the most extensively investigated has been the habituation of the orienting response to a repeated auditory stimulus. At first sight it may seem surprising that a habituation rather than an extinction model of desensitization should be under discussion (Evans, 1973), and the first task of a review of the habituation model must be to deal with this question.

If habituation applies to unconditioned responses and extinction to conditioned responses, the question of whether desensitization should be termed a habituation or an extinction process can be settled by definition on the basis of whether the responses being modified in desensitization are conditioned or unconditioned. However, the approach taken here is an empirical one, to see what empirical differences exist between habituation and extinction and whether the decremental processes that operate in desensitization are more closely analogous to

Requests for reprints should be sent to Fraser N. Watts, Department of Clinical Psychology, King's College Hospital, Denmark Hill, London SE5 9RS, England.

those found in habituation or to those found in extinction.

The most widely canvassed empirical difference between the two is that habituation is a short-term or temporary change in responsiveness, whereas extinction is a long-term or permanent one. Van Egeren (1971) has even suggested that this should be the defining distinction between habituation and extinction. However, such a proposal involves attaching a new meaning to the term habituation in view of the fact, acknowledged by Van Egeren, that the decrement of the orienting response on repetition of a stimulus can sometimes be relatively long-term. Though habituation may often be a relatively short-term process, it seems that unconditioned responses do not always show shorter term decrements than conditioned ones. Kimmel (1973), in a discussion of the differences between habituation and conditioning, reached the parallel conclusion that "habituation cannot be differentiated from conditioning simply on grounds of its temporariness" (p. 229).

It has been suggested by Razran (1971, p. 30) that habituation, unlike extinction, is not affected by cognition. However, this view of habituation seems to be incorrect. The rate of habituation is certainly modified by instructions that affect the set with which subjects approach stimuli (e.g., Pendergrass & Kimmel, 1968).

Another possible difference between the processes of habituation and those of extinction is that extinction produces a stimulus with response-suppressant properties but habituation does not. The crucial test of response suppression (Rescorla, 1969) is the summation test. If a stimulus has acquired response-suppressant properties as a result of habituation or extinction, when it is subsequently combined with a separate conditioned stimulus, it should produce a smaller response than the conditioned stimulus would produce on its own. Reiss and Wagner (1972) have shown that by this test habituation does not give response-suppressant properties to the stimulus concerned. If it were well established that extinction does indeed give a stimulus such properties, there would be a good reason for regarding habituation and extinction as different processes. However, the evidence on this point (Gray, 1975, pp. 105–106; Rescorla, 1969) is weak and so fails to establish a critical difference between the processes of habituation and extinction.

A number of other similarities and differences between habituation and extinction have been investigated. It is known that stimulant drugs retard both processes (Hilgard & Marquis, 1940; Lynn, 1966). In addition, temporal massing of stimuli normally facilitates both processes (Kling & Stevenson, 1970). Among differences that have been reported, Kling and Stevenson found that extinction produced an initial increase in responsiveness, though habituation did not. However, until such differences are replicated and their theoretical significance clarified, they can hardly provide an adequate empirical basis for concluding that different processes are involved in habituation and extinction. It can thus be concluded that no reliable empirical differences have been established between extinction and habituation on which the question of which process operates in desensitization could be settled. Thus, at present, it seems reasonable to use either habituation or extinction as a model for desensitization.

However, there are some general theoretical grounds for suggesting that the habituation of orienting responses may prove a particularly useful analogue for the decrement of anxiety in desensitization. They depend on the argument advanced by Gray (1975, 1976) that both novel stimuli that elicit orienting responses and stimuli that have been paired with punishing stimuli and arouse anxiety activate what Gray described as the "behavioural inhibition" system. This system produces an inhibition of ongoing behavior and increases arousal (though it does not produce the kind of response-suppressant effect measured by the summation test). Sedative drugs decrease the level of operation of the system. It is also relevant to note that this behavioral inhibition system has been found to be particularly active in neurotic introverts (Nicholson & Gray,

1972), who are also the personality group of which phobics are largely composed (Marks, 1969). So if it is correct that aversive stimuli and novel stimuli activate the same physiological–behavioral system, it seems reasonable to suggest that the processes of response decrement to both kinds of stimuli are similar. There is thus some reason for thinking that the habituation of orienting responses is worth pursuing as an analogue of response decrement in systematic desensitization.

Maximal Habituation Theory

The clearest available statement of the habituation model of desensitization is the maximal habituation hypothesis proposed by Lader and his colleagues (Lader & Mathews, 1968; Lader & Wing, 1966). The basic postulates of the model are that the rate of observed reduction in the magnitude of the galvanic skin response (GSR) to repeated presentations of phobic situations is a habituation process and that this habituation process is maximized by aspects of the procedure (notably relaxation) that lower central arousal (measured in terms of spontaneous GSR fluctuations). The model focuses entirely on the rate of response decrement as the dependent variable (measured as the slope of the linear decremental trend corrected for initial response levels). It is assumed that increasing the rate of response decrement in desensitization makes it a more effective treatment in the long term.

Individual Differences

The analogy between desensitization and habituation is supported by data on individual differences. Lader, Gelder, & Marks (1967) found a significant (.49) correlation across subjects between the rate of auditory habituation and clinical response to desensitization. In addition Lader (1967) showed that specific phobics exhibited both faster auditory habituation and a better clinical response to desensitization than other phobics.

These initial reports were followed by several similar studies. The most directly comparable study is that reported by Gillan and Rachman (1974), which failed to find a significant correlation between the same two measures. However, the number of subjects who received desensitization (16) was too small for this result to be given much weight. In addition, the correlation may have been lowered by the use of an experimental design that necessitated the omission of relaxation from the desensitization of half these subjects. Klorman (1974) also found a nonsignificant correlation between the rate of GSR decrement in auditory habituation and fear change after exposure to films of phobic stimuli, though the pattern of exposure of these films did not follow standard desensitization practice. Lang (1970), using a similar film exposure procedure in a complicated and somewhat incompletely reported experiment, has provided evidence that supports Lader et al.'s original finding. The rate of GSR habituation to tones of 50 dB and 100 dB correlated significantly with fear change as a result of the standard form of desensitization that includes relaxation. (The correlations were apparently low or insignificant when relaxation was omitted.) A number of significant correlations were also found between rate of GSR decrement to auditory tones and rate of decrement to phobic stimuli (Lang, 1970, p. 153), though there were too many insignificant correlations for the results as a whole to be convincing. In addition, the interpretation of this study is seriously affected by the fact that the rate of decrement to phobic stimuli did not correlate significantly with treatment outcome. This means that the relationship between treatment outcome and habituation rate to auditory stimuli was not mediated by the rate of anxiety decrement in desensitization. Instead of habituation rate directly influencing treatment outcome, it is more likely that both were a product of some other variable such as general arousal and that the specific mechanism relating arousal to outcome is something other than habituation rate.

All the above studies were concerned with GSR habituation, though heart-rate decrement has also been examined. Lang, Melamed and Hart (1970) reported a very high

correlation (.91) between heart rate decrement during desensitization and treatment outcome. However, Van Egeren (1970) found that the correlation between habituation rate to phobic versus neutral stimuli held only for skin conductance and not for heart rate and other variables. The explanation of such discrepancies between outcome indices can partly be explained in terms of their differential sensitivity at different arousal levels. As Lader (1975) has pointed out, the GSR is more sensitive at lower and heart rate at higher levels of arousal.

So far the support for the habituation model from studies of individual differences has been weak and inconsistent. However, the results should not be taken as counting strongly against the model. The proposition under consideration is that there are similar response decrement processes at work in auditory habituation and systematic desensitization. It seems to be a general feature of psychology that outcome measures in two different response systems can be the result of similar processes and that these measures correlate only weakly across subjects. Shapiro (1966) has put this forward as a general proposition, giving as one example that there can be low correlations across subjects between the strengths of different conditioned responses (GSR, eyeblink, etc.), but nevertheless in both cases the conditioned responses are affected by the same variables, such as frequency of reinforcement. In the same way, response decrement in auditory habituation and systematic desensitization could be the result of similar processes, without the degree of decrement found in these two situations necessarily being highly correlated across individuals.

Relaxation

Reciprocal inhibition theory (Wolpe, 1958) sees the role of relaxation training in desensitization as producing a state of muscular relaxation incompatible with the state of anxiety usually elicited by phobic stimuli. This state of reciprocal inhibition is turned, by reinforcement, into a more permanent state of conditioned inhibition. In contrast, the maximal habituation model (Lader & Mathews, 1968) views the role of relaxation as lowering central arousal (measured by spontaneous GSR fluctuations). This in turn increases the rate of response decrement.

In support of this view, Mathews and Gelder (1969) showed that relaxation does indeed reduce the frequency of spontaneous GSR fluctuations. In addition, Lader and Wing (1966) showed that sedatives had a similar effect and also resulted in a faster rate of GSR decrement to auditory stimuli. It thus seems reasonable to suppose that the arousal-lowering effect of relaxation has a similar effect on the rate of response decrement in systematic desensitization. Subsequent evidence has generally supported the view that relaxation lowers arousal, though relaxation is apparently not a uniquely effective procedure for achieving this result. The results of experiments that examine its effects therefore depend on what control procedure relaxation is compared with. Relaxation produces lower levels of arousal than attention to instructions (Mathews & Gelder, 1969), than an "eyes-open" control condition (Teasdale, 1971), or than looking at a neutral slide (Benjamin, Marks, & Huson, 1972), but is not more effective than procedures that simply omit Jacobsonian muscular relaxation exercises but that would in other respects be equally likely to induce a state of lowered arousal (Edelman, 1971; Grossberg, Note 1).

Support for the assumption that relaxation increases the rate of response decrement in desensitization has been weaker. There is some supporting evidence from Van Egeren (1971) and Wolpe and Flood (1970), though others (Benjamin et al., 1972; Waters & McDonald, 1973; Waters, McDonald, & Koresko, 1972) have reported negative results. One of the problems in interpreting some of these results is that relaxation seems to result in increased responsiveness to the initial presentation of a phobic stimulus. For example, that Wolpe and Flood (1970) reported response decrement with relaxation, but not without it, seems to be entirely attributable to the larger initial responses in the relaxation condition. There is also some

doubt about whether relaxation produces faster decrement of subjective anxiety. Waters et al. (1972) reported that subjects reached the criterion of no subjective anxiety in fewer trials with relaxation, but Benjamin et al. (1972) found no differences in subjective anxiety decrement between a relaxed and a nonrelaxed condition.

It is even more doubtful whether relaxation facilitates response decrement in auditory habituation. Lader and Mathews (1968) had no relevant evidence at the time they published their theory. However, this was subsequently investigated by Teasdale (1971) in a series of four experiments, using tones of 70 dB and 92 dB. He found no effect of relaxation on the rate of response decrement. In addition, Freeling (1972) reported that relaxation did not increase response decrement to a series of pistol shots. The assumption that relaxation increases the rate of response decrement has thus received only very weak support, and it is doubtful whether a version of the habituation model that has the effect of relaxation as a central assumption can be maintained. However, an alternative version of the habituation model of desensitization is elaborated in the remainder of this article, which does not assume a direct effect of relaxation on the habituation process.

Dual-Process Habituation Theory

So far the term *habituation* has been used to refer to an observable process of response decrement, but the recent development of dual-process habituation theory (Groves & Thompson, 1970; Thompson, Groves, Teyler, & Roemer, 1973) makes it necessary to clarify this usage. Dual-process theory proposes that observable response decrement is the summation of two inferred processes; habituation and sensitization. Response increment is equally the summation of the same two inferred processes.

It is not obvious that sensitization is a good term for the process of incremental change in responsiveness to a repeated unconditioned stimulus that Groves and Thompson referred to. The term already has a different technical meaning in the context of conditioning theory (e.g., Kimble, 1961), but it seems that even more confusion would be caused by adopting a different term in the present article.

Habituation and sensitization are most conveniently defined by contrasting their characteristics:

1. Habituation is a purely decremental process, whereas sensitization at first grows and then decays.
2. Habituation affects a particular response to a particular stimulus, whereas sensitization affects general responsiveness.
3. Habituation is independent of stimulus intensity (Thompson et al., 1973), whereas sensitization is positively related to stimulus intensity.
4. Though both habituation and sensitization decay spontaneously, sensitization is the more transient phenomenon (Davis, 1972).
5. Repeated series of habituation training trials result in progressively more habituation but in progressively less sensitization.

Groves and Thompson's theory resembles the conclusions reached by Hinde (1966) about change in responsiveness on the basis of his review of the relevant ethological research. Hinde also found it necessary to postulate an incremental process that accompanies the decremental one. There is also agreement that the decremental process is relatively stimulus specific and that the incremental process is a generalized one.

It is clear that response increment can result from exposure to a phobic stimulus. For example, Miller and Levis (1971) gave snake-phobic subjects two avoidance tests separated by a 50-minute interval. There were four experimental groups who spent respectively 0, 15, 30, and 45 minutes of this time in visual observation of the snake. The no-exposure group showed the least avoidance behavior during the second test (and significantly less avoidance than the 15-minute-exposure group). This indicates that exposure activated an incremental process. For the three groups that received some degree of exposure to the snake between tests, there was a tendency for the amount of exposure to be negatively associated with

avoidance behavior during the second test. Stone and Borkovec (1975) obtained similar results. The no-snake-exposure and the 45-minute-snake-exposure groups showed less avoidance behavior and autonomic arousal during the postexposure test than did the 15-minute-exposure group. The greater responsivenes of the 15-minute than the no-exposure group during the posttest indicates the presence of some kind of incremental process, whereas that this increased responsiveness was not found in the 45-minute-exposure group is consistent with the decay of sensitization over this period.

Relaxation

The maximal habituation theory (Lader & Mathews, 1968) proposes that the role of relaxation in desensitization is to lower arousal and that this in turn facilitates the rate of habituation. Dual-process theory invites a different view, namely, that relaxation reduces the amount of sensitization that takes place. Relaxation may not affect habituation at all, but its effects on sensitization summate with the habituation process to inrease the rate of response decrement.

As sensitization is a relatively transient process, the facilitatory effect of relaxation on response decrement would be only short-term. It is relevant in this connection to note that Teasdale (1971) found that relaxation had no immediate effect on auditory habituation, but resulted in less long-term decrement. Whether relaxation reduces the amount of long-term anxiety reduction in desensitization also is not clear. One problem in examining this is that an avoidance test given immediately after treatment would affect performance during a subsequent avoidance test. Subjects who received desensitization with relaxation might do better at follow-up simply because they had done better at an immediate posttreatment test and this had had an anxiety-reducing effect. It would therefore be necessary to use a design in which different subjects were used for testing the immediate and the delayed effects of relaxation on response decrement in desensitization.

Stimulus Intensity

It has been shown that the use of a graded incremental stimulus hierarchy facilitates response decrement to repeated auditory stimuli (Davis & Wagner, 1969; Groves & Thompson, 1970), and this has been attributed, in the context of dual-process theory, to the minimization of sensitization rather than to the maximization of habituation. Dual-process theory also proposes that the value of a graded hierarchy in desensitization, like that of relaxation, is to minimize sensitization. If this is correct, it implies, as Davis (1972) pointed out, that the beneficial effects of an incremental hierarchy would be relatively transient and might not appear on delayed testing. That Krapfl and Nawas (1970) found that the use of a graded hierarchy in desensitization was no more beneficial during testing immediately after treatment than during the follow-up counts against this view. However, the superiority of a graded hierarchy in the short term is more likely to appear with severely handicapped phobics. Krapfl and Nawas's experiment, like many others, used student snake phobics. In addition, as with the effects of relaxation, the immediate and long-term effect of a graded hierarchy should be tested in different subjects. Two other experiments (Klorman, 1974; Lang, 1970) that also failed to find a short-term advantage in an incremental hierarchy cannot be given much weight, as they used only two or three different intensity levels. Davis and Wagner (1969) showed for auditory habituation that a hierarchy is only helpful if many small incremental steps are used.

It is important in discussing the effects of stimulus intensity to make the distinction between *relative* and *absolute* habituation (Davis & Wagner, 1968). Relative habituation refers to a situation in which habituation is produced and tested with the same stimulus. Absolute habituation refers to a situation in which habituation to one stimulus is tested with another stimulus. The distinction is especially important if the test stimulus is more intense than that used for habituation training. The conditions that are most favorable for absolute and relative ha-

bituation are not necessarily the same. In particular, Davis and Wagner (1968) showed that high-intensity stimuli resulted in more absolute but not in more relative habituation at subsequent testing. Thus, an incremental hierarchy in desensitization might facilitate initial response decrement to the stimuli used in treatment without increasing the amount of response decrement to a separate set of test stimuli. The clinical objective of desensitization is to reduce responsiveness to stimuli of varying intensities that occur in the natural environment and not just to reduce responsiveness to the stimuli used in desensitization training. The potential advantage of habituation to high-intensity stimuli (i.e., flooding) is that it may produce more response decrement when assessed by these "absolute" standards. This assumes, of course, that the sensitization produced by flooding can be successfully managed.

Dual-process habituation theory is certainly better able than maximal habituation theory to explain that both desensitization and flooding are effective anxiety-reduction procedures. It is a serious paradox for maximal habituation theory that a treatment that uses conditions (prolonged exposure to high-intensity stimuli) resulting in high levels of arousal should achieve response decrement at all. The explanation, in terms of dual-process theory, is that whereas desensitization minimizes sensitization, flooding elicits it but also provides long enough exposures for it to decay again. Long sessions are important in getting the best results from flooding (e.g., Stern & Marks, 1973). The repeated growth and decay of sensitization should reduce the extent to which sensitization occurs in the future. One implication of this is that flooding should produce a more generalized reduction in responsiveness than should desensitization, and consistent with this prediction is that Watson and Marks (1971) have shown that flooding to relevant and irrelevant fears is equally effective in the treatment of phobics. On the other hand, it is predicted that systematic desensitization (which, because it uses low-intensity stimuli, is presumably based on habituation rather than on the decay of sensitization) would be markedly more effective if relevant rather than irrelevant fears were used.

Stimulus Lengths

Desensitization usually employs short (e.g., 10 sec) presentations of imaginal stimuli, and normally presentations are terminated sooner if anxiety is reported. Dual-process theory sees the role of short stimulus presentations as also preventing sensitization. They would be especially helpful in doing this if other conditions (i.e., relaxation and low-intensity stimuli) were such that they minimized sensitization. If relaxation and low-intensity stimuli were not used, short presentations of stimuli would probably be able to do little to prevent sensitization from developing, and longer presentations, which allow more time for inferred habituation to take place, would result in a greater degree of response decrement.

There are two experiments that have found such an interaction effect between stimulus lengths and relaxation. Proctor (1969) found that when relaxation was used, desensitization with 5-sec exposures to slides produced more change in subjective anxiety during a subsequent avoidance test than did 20-sec exposures. The reverse obtained if relaxation was omitted. Sue (1975) reported a similar interaction in comparing 5-sec and 30-sec stimulus lengths. In desensitization without relaxation, longer stimulus exposures produced more change on a behavioral avoidance test and on the Fear Survey Schedule (Geer, 1965), but had no significant effect when relaxation was not used. Watts (1971) reported a similar interaction between stimulus intensity and stimulus length. Following a suggestion made by Koepke and Pribram (1966) in their attempt at resolving some apparently conflicting findings in the habituation literature, Watts predicted and found that relatively low-intensity desensitization items habituated to zero anxiety more rapidly with short (5 sec) than with long (30 sec) presentations but that the reverse obtained for higher intensity items.

The advantage of longer presentations is best attributed to the allowance of more time

for habituation to take place rather than to the decay of sensitization. The time scale (30 sec) is probably too short for the decay of sensitization. But in any case, the advantage of longer stimulus presentations is known to be a relatively long-term one, whereas sensitization is relatively transitory in its effects.

Watts (1971) provided evidence for the long-term value of longer stimulus presentation in desensitization. He found that even when reduction to zero anxiety was achieved more rapidly with short presentations, more long-term desensitization was achieved with longer presentations. In the short term the combination of short presentations of low-intensity stimuli and relaxation can apparently prevent the development of sensitization and so appear to facilitate response decrement, but little long-term advantage results from this. Where relaxation or low-intensity stimuli are not used, short presentations are apparently not sufficient to prevent the accumulation of sensitization, and there is not even any short-term advantage in using short stimulus presentations. In this case there is both an immediate and a long-term advantage in using longer stimulus presentations.

Watts (1974) considered in more detail the mechanism by which longer stimulus presentations result in more long-term habituation. He suggested that the amount of long-term response decrement depends on the extent to which the subject forms a clear model of the stimulus (c.f. Sokolov, 1963). In support of this hypothesis, Watts was able to show that describing desensitization items each time they were presented resulted in more long-term anxiety reduction, though it did not affect the rate of response decrement.

It may well be that some of the confusion in the literature about whether relaxation facilitates desensitization could be resolved by taking lengths of stimulus presentation into account. Crowder and Thornton (1970) have drawn attention to the fact that studies such as their own that failed to show that relaxation facilitates desensitization tended to use relatively long presentations. Similarly, it is predicted that the facilitatory effects of a graded stimulus hierarchy would be easier to demonstrate if short stimulus presentations were used.

The interaction effects that have been found between stimulus length and relaxation and between stimulus length and stimulus intensity prove a problem for most theories of desensitization. There have so far been no suggestions as to how they could be predicted from reciprocal inhibition theory or maximal habituation theory. Equally, cognitive theories of desensitization (Kazdin & Wilcoxon, 1976) would have no basis for predicting these effects. They thus provide important empirical support for the dual-process model of desensitization, with which they are in accord.

Interstimulus Interval Length

Another example of the differential effects of procedural variables on short-term and long-term change concerns interstimulus intervals. Though short intervals usually facilitate initial response habituation, they do not increase the amount of long-term habituation (Askew, 1970; Davis, 1970). This facilitatory effect of short-term intervals can be attributed to refractory-period effects. Watts (1973) found that in desensitization also, interstimulus interval lengths have a short-term but not a long-term effect. However, the short-term effect is different from that found in auditory habituation. Longer intervals resulted in faster initial response decrement to relatively high-intensity desensitization items, though they had no effect with low-intensity items. Presumably the longer intervals prevented the accumulation of the sensitization that can occur with high-intensity stimuli, but this preventative effect would not be expected to have any long-term benefit.

Long-Term Anxiety Reduction

One of the central weaknesses of the maximal habituation model, as developed by Lader and Mathews (1968), is its failure to take account of the long-term effects of procedural variables. In particular, as has been argued, Lader and Mathews were probably

incorrect to assume that aspects of the desensitization procedure, such as relaxation and an incremental stimulus hierarchy, that appear to increase the initial rate of anxiety decrement in the short term are also of long-term benefit. In addition, short stimulus lengths and short interstimulus intervals can facilitate anxiety reduction in the short term, but can be of no long-term benefit. In the context of the dual-process theory elaborated here, it has been suggested that variables that facilitate short-term but not long-term decrement operate on sensitization rather than on habituation. But whether or not this is correct, it is now clear that theories of desensitization that do not give explicit consideration to both immediate and long-term effects cannot be adequate. There is a need for a great deal more research on the delayed effects of desensitization sessions. The few investigations published so far (e.g., Agras, 1965; Rachman, 1966) leave many questions unanswered. It is also necessary to become more precise about what is meant by *long-term* effects. Testing at delays of 1 hour, 24 hours, and 1 week may produce quite different results, though it would be impossible to tease out such differences from the currently available research. In the meantime, the dual-process theory developed here has the unique distinction of making specific predictions about immediate and long-term effects.

Conclusion

It is probably too soon to make a final judgment on the adequacy of the habituation model of desensitization, though it can at least be claimed that it is a viable alternative to the reciprocal inhibition theory. Moreover, it has served the useful function of generating fresh problems and hypotheses for investigation. In particular, it has tended to generate more detailed research on the pattern of anxiety reduction than have other theories. In this way it has resulted in some significant advances in our understanding of the processes that operate in systematic desensitization.

Some data have already accumulated (e.g., the interaction effects between stimulus length and relaxation/stimulus intensity) that can be predicted from the dual-process theory, but that appear to pose problems for any other current theory. The need for the future is to make the dual-process theory increasingly precise so that specific predictions from the theory can be tested and, if necessary, refuted. This article has tried to make a contribution to this theoretical task.

Reference Note

1. Grossberg, J. M. *The physiological effectiveness of brief training in differential muscle relaxation* (Tech. Rep. 9). La Jolla, Calif.: Western Behavioral Sciences, 1965.

References

Agras, S. An investigation of decrements of anxiety responses during systematic desensitization. *Behaviour Research and Therapy*, 1965, *2*, 267–270.

Askew, H. R. Effects of stimulus intensity and intertrial interval on habituation of the head-shake response in the rat. *Journal of Comparative and Physiological Psychology*, 1970, *72*, 492–497.

Benjamin, S., Marks, I. M., & Huson, J. Active muscular relaxation in desensitization of phobic patients. *Psychological Medicine*, 1972, *2*, 381–390.

Crowder, J. E., & Thornton, D. W. Effects of systematic desensitization, programmed fantasy and bibiotherapy on a specific fear. *Behaviour Research and Therapy*, 1970, *8*, 35–41.

Davis, M. Effects of interstimulus interval lengths and variability on startle-response habituation in the rat. *Journal of Comparative and Physiological Psychology*, 1970, *72*, 177–192.

Davis, M. Differential retention of sensitization and habituation of the startle response in the rat. *Journal of Comparative and Physiological Psychology*, 1972, *78*, 260–267.

Davis, M., & Wagner, S. R. Startle responsiveness after habituation to different intensities of tone. *Psychonomic Science*, 1968, *12*, 337–338.

Davis, M., & Wagner, A. R. Habituation of startle response under incremental sequence of stimulus intensities. *Journal of Comparative and Physiological Psychology*, 1969, *67*, 486–492.

Edelman, R. I. Desensitization and physiological arousal. *Journal of Personality and Social Psychology*, 1971, *17*, 259–266.

Evans, I. The logical requirements for explanations of systematic desensitization. *Behavior Therapy*, 1973, *4*, 506–514.

Freeling, N. W. The psychophysiological effects of brief relaxation training: A test of the maximal habituation hypothesis (Doctoral dissertation, Bowling Green State University, 1971). *Dissertation Abstracts International*, 1972, *32*, 4856B–4857B. (University Microfilms No. 72-5289)

Geer, J. H. The development of a scale to measure fear. *Behaviour Research and Therapy,* 1965, *3,* 45–53.

Gillan, P., & Rachman, S. An experimental investigation of desensitization in phobic patients. *British Journal of Psychiatry,* 1974, *124,* 392–401.

Gray, J. A. *Elements of a two-process theory of learning.* London: Academic Press, 1975.

Gray, J. A. The behavioural inhibition system: A possible substrate for anxiety. In M. P. Feldman & A. Broadhurst (Eds.), *Theoretical and experimental bases of the behaviour therapies.* London: Wiley, 1976.

Groves, P. M., & Thompson, R. F. Habituation: A dual-process theory. *Psychological Review,* 1970, *77,* 419–459.

Hilgard, E. R., & Marquis, D. G. *Conditioning and learning.* New York: Appleton-Century-Crofts, 1940.

Hinde, R. A. *Animal behaviour: A synthesis of ethology and comparative psychology.* New York: McGraw-Hill, 1966.

Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin,* 1976, *83,* 729–759.

Kimble, G. A. (Ed.). *Hilgard and Marquis' Conditioning and learning.* London: Methuen, 1961.

Kimmel, H. D. Habituation, habituality and conditioning. In H. V. S. Peeke & M. J. Herz (Eds.), *Habituation* (Vol. 1). New York: Academic Press, 1973.

Kling, J. W., & Stevenson, J. G. Habituation and extinction. In G. Horn & R. A. Hinde (Eds.), *Short-term changes in neural activity and behaviour.* Cambridge: Cambridge University Press, 1970.

Klorman, R. Habituation of fear: Effects of intensity and stimulus order. *Psychophysiology,* 1974, *11,* 15–26.

Koepke, J. E., & Pribram, K. H. Habituation of the GSR as a function of stimulus duration and spontaneous activity. *Journal of Comparative and Physiological Psychology,* 1966, *61,* 442–448.

Krapfl, J. E., & Nawas, M. M. Differential ordering of stimulus presentation in systematic desensitization. *Journal of Abnormal Psychology,* 1970, *75,* 333–337.

Lader, M. H. Palmar conductance measures in anxiety and phobic states. *Journal of Psychosomatic Research,* 1967, *11,* 271–281.

Lader, M. H. *The psychophysiology of mental illness.* London: Routledge & Kegan Paul, 1975.

Lader, M. H., Gelder, M. G., & Marks, I. M. Palmar skin-conductance measures as predictors of response to desensitization. *Journal of Psychosomatic Research,* 1967, *11,* 283–290.

Lader, M. H., & Mathews, A. M. A physiological model of phobic anxiety and desensitization. *Behavior Research and Therapy,* 1968, *6,* 411–421.

Lader, M. H., & Wing, L. *Physiological measures, sedative drugs and morbid anxiety.* London: Oxford University Press, 1966.

Lang, P. J. Stimulus control, response control and the desensitization of fear. In D. J. Levis (Ed.), *Learning approach to therapeutic behavior change.* Chicago: Aldine, 1970.

Lang, P. J., Melamed, B. G., & Hart, J. A psychophysiological analysis of fear modification using an automated desensitization procedure. *Journal of Abnormal Psychology,* 1970, *76,* 220–234.

Lynn, R. *Attention, arousal and the orientation reaction.* Oxford: Pergamon Press, 1966.

Marks, I. M. *Fears and phobias.* London: Academic Press, 1969.

Mathews, A. M., & Gelder, M. G. Psychophysiological investigation of brief relaxation training. *Journal of Psychosomatic Research,* 1969, *13,* 1–12.

Miller, B. V., & Levis, D. J. The effects of varying short visual exposure times to a phobic stimulus on subsequent avoidance behaviour. *Behaviour Research and Therapy,* 1971, *9,* 17–21.

Nicholson, J. N., & Gray, J. A. Peak shift, behavioural contrast and stimulus generalization as related to personality and development in children. *British Journal of Psychology,* 1972, *63,* 47–62.

Pendergrass, V. E., & Kimmel, H. D. UCR diminution in temporal conditioning and habituation. *Journal of Experimental Psychology,* 1968, *77,* 1–6.

Proctor, S. Duration of exposure to items and pretreatment training as factors in systematic desensitization therapy. In R. D. Rubin & C. M. Franks (Eds.), *Advances in behavior therapy, 1968.* New York: Academic Press, 1969.

Rachman, S. Studies in desensitization: III. Speed of generation. *Behaviour Research and Therapy,* 1966, *4,* 7–15.

Razran, G. H. S. *Mind in evolution: An East-West synthesis of learned behavior and cognition.* Boston: Houghton-Mifflin, 1971.

Rescorla, R. A. Pavlovian conditioned inhibition. *Psychological Bulletin,* 1969, *72,* 77–94.

Reiss, S., & Wagner, A. R. CS habituation produces a "latent inhibition effect" but no active "conditioned inhibition." *Learning and Motivation,* 1972, *3,* 237–245.

Shapiro, M. B. Generality of psychological processes and specificity of outcomes. *Perceptual and Motor Skills,* 1966, *23,* 16.

Sokolov, Y. N. *Perception and the conditional reflex.* Oxford: Pergamon Press, 1963.

Stern, R., & Marks, I. Brief and prolonged flooding: A comparison in agoraphobic patients. *Archives of General Psychiatry,* 1973, *28,* 270–276.

Stone, N. M., & Borkovec, T. D. The paradoxical effect of brief CS exposure on analogue phobic subjects. *Behaviour Research and Therapy,* 1975, *13,* 51–54.

Sue, D. The effect of duration of exposure on systematic desensitization and extinction. *Behaviour Research and Therapy*, 1975, *13*, 55–60.

Teasdale, J. D. *Relaxation and habituation: An experimental and theoretical investigation.* Unpublished doctoral dissertation, London University, 1971.

Thompson, R. F., Groves, P. M., Teyler, T. J., & Roemer, R. A. A dual-process theory of habituation: Theory and behavior. H. V. S. Peeke & M. J. Herz (Eds.), *Habituation* (Vol. 1). New York: Academic Press, 1973.

Van Egeren, L. F. Psychophysiology of systematic desensitization: The habituation model. *Journal of Behavior Therapy and Experimental Psychiatry*, 1970, *1*, 249–255.

Van Egeren, L. F. Psychophysiological aspects of systematic desensitization: Some outstanding issues. *Behaviour Research and Therapy*, 1971, *9*, 65–77.

Waters, W. F., & McDonald, D. G. Autonomic response to auditory, visual and imagined stimuli in a systematic desensitization context. *Behaviour Research and Therapy*, 1973, *11*, 577–585.

Waters, W. F., McDonald, D. G., & Koresko, R. L. Psychophysiological responses during analogue systematic desensitization. *Behaviour Research and Therapy*, 1972, *10*, 381–393.

Watson, J. P., & Marks, I. M. Relevant and irrelevant fear in flooding: A crossover study of phobic patients. *Behavior Therapy*, 1971, *2*, 275–293.

Watts, F. N. Desensitization as an habituation phenomenon: I. Stimulus intensity as a determinant of the effects of stimulus lengths. *Behaviour Research and Therapy*, 1971, *9*, 209–217.

Watts, F. N. Desensitization as an habituation phenomenon: II. Studies of interstimulus interval length. *Psychological Reports*, 1973, *33*, 715–718.

Watts, F. N. The control of spontaneous recovery of anxiety in imaginal desensitization. *Behaviour Research and Therapy*, 1974, *12*, 57–59.

Wolpe, J. *Psychotherapy by reciprocal inhibition.* Stanford: Stanford University Press, 1958.

Wolpe, J., & Flood, J. The effect of relaxation on the galvanic skin response to repeated phobic stimuli in ascending order. *Journal of Behavior Therapy and Experimental Psychiatry*, 1970, *1*, 195–200.

Received February 14, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 3, 638–641

The "File Drawer Problem" and Tolerance for Null Results

Robert Rosenthal
Harvard University

For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results. Quantitative procedures for computing the tolerance for filed and future null results are reported and illustrated, and the implications are discussed.

Both behavioral researchers and statisticians have long suspected that the studies published in the behavioral sciences are a biased sample of the studies that are actually carried out (Bakan, 1967; McNemar, 1960; Smart, 1964; Sterling, 1959). The extreme view of this problem, the "file drawer problem," is that the journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g., $p > .05$) results.

In the past there was very little one could do to assess the net effect of studies, tucked away in file drawers, that did not make the magic .05 level (Rosenthal & Gaito, 1963, 1964). Now, however, although no definitive solution to the problem is available, one can establish reasonable boundaries on the problem and estimate the degree of damage to any research conclusion that could be done by the file drawer problem.

This advance in our ability to cope with the file drawer is an outgrowth of the increasing interest of behavioral scientists in summarizing bodies of research literature systematically and quantitatively, both with respect to significance levels (Rosenthal, 1969, 1976, 1978) and with respect to effect-size estimation (Hall, 1978; Rosenthal, 1969, 1976; Rosenthal & Rosnow, 1975; Smith & Glass, 1977; Glass, Note 1). One hopes that this interest in summarizing entire research domains will lead to an improvement in bookkeeping so that eventually all results will be recorded both with an estimate of effect size (e.g., r or d; Cohen, 1977) and with the level of significance obtained, or more practically, with the standard normal deviate (Z) that corresponds to the obtained p (Rosenthal, 1978).[1] Future appraisals of research domains of the type found in *Psychological Bulletin* should give estimates of overall effect sizes and significance levels; these estimates of overall significance can provide a basis for coping with the file drawer problem.

Tolerance for Future Null Results

Given any systematic quantitative review of the literature bearing on a particular hy-

Preparation of this article was supported in part by the National Science Foundation.

I would like to thank Judith A. Hall and Donald B. Rubin for their valuable improvements of an earlier version of this article.

Requests for reprints should be sent to Robert Rosenthal, Department of Psychology and Social Relations, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138.

[1] Standard normal deviates (Z) can be found by various methods, of which the following three are most often useful: (*a*) Obtain the exact p associated with the test statistic (e.g., t, F, or χ^2) and find the Z associated with that p in tables of the normal distribution; (*b*) if the effect size r or phi is given or can be computed, Z can be estimated by $r(N)^{\frac{1}{2}}$; (c) if the effect size d is given or can be computed, Z can be estimated by $[d^2/(d^2+4)]^{\frac{1}{2}}(N)^{\frac{1}{2}}$.

pothesis, for example, that psychotherapy is effective (Glass, Note 1), that women are more sensitive than men to nonverbal cues (Hall, 1978), or that one person's expectation for another person's behavior can come to serve as self-fulfilling prophecy (Rosenthal, 1969, 1976), it is easy to calculate an overall probability, based on all the independent studies available to the reviewer, that the effect in question is "real," that is, not a Type I error (Rosenthal, 1978). The fundamental idea in coping with the file drawer problem is simply to calculate the number of studies averaging null results that must be in the file drawers before the overall probability of a Type I error is brought to any desired level of significance, say, $p = .05$. This number of filed studies, or the tolerance for future null results, is then evaluated for whether such a tolerance level is small enough to threaten the overall conclusion drawn by the reviewer. If the overall level of significance of the research review will be brought down to the level of *just significant* by the addition of just a few more null results, the finding is not resistant to the file drawer threat.

Computation

Perhaps the simplest, most useful way of computing the overall p of a set of research studies is the method of adding Zs (Cochran, 1954; Mosteller & Bush, 1954; Rosenthal, 1978). This method requires only that one add the standard normal deviates of Zs associated with the ps obtained and divide by the square root of the number of studies being combined. The result is itself a Z that can be entered in a table to find the associated overall p:

$$Z_c = k\bar{Z}_k/\sqrt{k} = \sqrt{k}\bar{Z}_k, \qquad (1)$$

where Z_c is the new combined Z, k is the number of studies combined, and $\bar{Z}_k$ is the mean Z obtained for the k studies.

To find the number (X) of new, filed, or unretrieved studies averaging null results required to bring the new overall p to any desired level, say, just significant at $p = .05$ ($Z = 1.645$), one simply writes:

$$1.645 = k\bar{Z}_k/\sqrt{k + X}. \qquad (2)$$

Rearrangement shows, then, that

$$X = (k/2.706)[k(\bar{Z}_k)^2 - 2.706]. \qquad (3)$$

An alternative formula that may be more convenient when the sum of the Zs (ΣZ) is given rather than the mean Z is as follows: $X = [(\Sigma Z)^2 / 2.706] - k$. One method based on counting rather than adding Zs may be easier to compute and can be employed when exact p levels are not available; but it is probably less powerful. If X is the number of new studies required to bring the overall p to .50 (not to .05), s is the number of summarized studies significant at $p < .05$, and n is the number of summarized studies not significant at .05, then $X = 19s - n$. Another conservative alternative when exact p levels are not available is to set $Z = .00$ for any nonsignificant result and to set $Z = 1.645$ for any result significant at $p \leq .05$.

Equations 1, 2, and 3 all assume that each of the k studies is independent of all other $k - 1$ studies, at least in the sense of employing different sampling units. There are other senses of independence, however; for example, one can think of two or more studies conducted in a given laboratory as less independent than two or more studies conducted in different laboratories. Such nonindependence can be assessed by intraclass correlations. Whether nonindependence of this type serves to increase Type I or Type II errors appears to depend in part on the relative magnitude of the Zs obtained from the studies that are correlated or too similar. If the correlated Zs are, on the average, as high (or higher) as the grand mean Z corrected for nonindependence, the combined Z one computes by treating all studies as independent will be too large. If the correlated Zs are, on the average, clearly low relative to the grand mean Z corrected for nonindependence, the combined Z one computes by treating all studies as independent will tend to be too small.

Illustration

In 1969, 94 experiments examining the effects of interpersonal self-fulfilling prophecies were summarized (Rosenthal, 1969). The mean Z of these studies was 1.014, k was 94, and Z_c for the studies combined was $9.83 = 94(1.014)/(94)^{\frac{1}{2}}$.

How many new, filed, or unretrieved studies (X) would be required to bring this very large Z down to a barely significant level ($Z = 1.645$)? By Equation 3,

$$X = (94/2.706)\ [94(1.014)^2 - 2.706] = 3{,}263.$$

One finds that 3,263 studies averaging null results ($\bar{Z} = .00$) must be crammed into file drawers before one would conclude that the overall results were due to sampling bias in the studies summarized by the reviewer. In a more recent summary of the same area of research (Rosenthal, 1976), the mean Z of 311 studies was 1.180, k was 311, and X was 49,457! Thus, nearly 50,000 unreported studies averaging a null result would have to exist somewhere before the overall results could reasonably be ascribed to sampling bias.

Discussion

There is both a sobering and a cheering lesson to be learned from careful study of Equation 3. The sobering lesson is that small numbers of studies that are not very significant, even when their combined p is significant, may well be misleading in that only a few studies filed away could change the combined significant result to a nonsignificant one. Thus, 15 studies averaging a Z of .50 have a combined p of .026; but if there were only 6 studies tucked away showing a mean Z of .00, the tolerance level for null results would be exceeded, and the significant result would become nonsignificant (i.e., $p > .05$). Or if there were 2 studies averaging a Z of 2.00, the combined p would be about .002; but uncovering 4 new studies averaging a Z of .00 would bring p into the *not significant* region.

The cheering lesson is that when the number of studies available grows large or the mean directional Z grows large, the file drawer hypothesis as a plausible rival hypothesis can be safely ruled out. If 300 studies are found to average a Z of $+1.00$, it would take 32,960 studies to bring the new combined p to a nonsignificant level; that many file drawers full is simply too improbable.

At the present time no firm guidelines can be given as to what constitutes an unlikely number of unretrieved or unpublished studies. For some areas of research 100 or even 500 unpublished and unretrieved studies may be a plausible state of affairs, whereas for others even 10 or 20 seems unlikely. Probably any rough and ready guide should be based partly on k so that as more studies are known it becomes more plausible that other studies in that area may be in those file drawers. Perhaps one could regard as resistant to the file drawer problem any combined results for which the tolerance level (X) reaches $5k + 10$. This seems a conservative but reasonable tolerance level; the $5k$ portion suggests that it is unlikely that the file drawers have more than five times as many studies as the reviewer, and the 10 sets the minimum number of studies that could be filed away at 15 (when $k = 1$).

It appears that more and more reviewers of research literature are estimating average effect sizes and combined ps of the studies they summarize. It would be very helpful to readers if for each combined p they presented, reviewers also gave the tolerance for future null results associated with their overall significance level.

Reference Note

1. Glass, G. V. *Primary, secondary, and meta-analysis of research*. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

References

Bakan, D. *On method*. San Francisco: Jossey-Bass, 1967.

Cochran, W. G. Some methods for strengthening the common χ^2 tests. *Biometrics*, 1954, *10*, 417–451.

Cohen, J. *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press, 1977.

Hall, J. A. Gender effects in decoding nonverbal cues. *Psychological Bulletin*, 1978, *85*, 845–857.

McNemar, Q. At random: Sense and nonsense. *American Psychologist*, 1960, *15*, 295–300.

Mosteller, F. M., & Bush, R. R. Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Vol. 1. Theory and method.* Cambridge, Mass.: Addison-Wesley, 1954.

Rosenthal, R. Interpersonal expectations. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press, 1969.

Rosenthal, R. *Experimenter effects in behavioral research* (Enlarged ed.). New York: Irvington, 1976.

Rosenthal, R. Combining results of independent studies. *Psychological Bulletin*, 1978, *85*, 185–193.

Rosenthal, R., & Gaito, J. The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 1963, *55*, 33–38.

Rosenthal, R., & Gaito, J. Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, 1964, *15*, 570.

Rosenthal, R., & Rosnow, R. L. *The volunteer subject*. New York: Wiley-Interscience, 1975.

Smart, R. G. The importance of negative results in psychological research. *Canadian Psychologist*, 1964, *5*, 225–232.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 1977, *32*, 752–760.

Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 1959, *54*, 30–34.

Received February 16, 1978 ■

CONTENTS (continued)

5 12/79

Ro
Joder 8
18 Ju 79
13/4/79

VOLUME 86 / NUMBER 4 JULY 1979

Psychological Bulletin

(Continued on inside back cover)

BLISHED BIMONTHLY BY THE AMERICAN PSYCHOLOGICAL ASSOCIATION
ISSN 0033-2909

The *Psychological Bulletin* publishes evaluative reviews and interpretations of substantive and methodological issues in the psychological research literature. The Journal reports original research only when it illustrates some methodological problem or issue. Discussions of methodological issues should be aimed at the solution of some particular research problem on psychology, but should be of sufficient breadth to interest a wide readership among psychologists; articles of a more specialized nature can be directed to the various statistical, psychometric, and methodological journals. The *Bulletin* does not publish original theoretical articles; these should be submitted to the *Psychological Review*.

Abstracts: All articles must be preceded by an abstract of 100–175 words. Detailed instructions for preparation of abstracts appear in the *Publication Manual of the American Psychological Association* (2nd ed.), or they may be obtained from the Editor or from APA Central Office.

Blind review: Because reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the author's identity or affiliation should be on separate pages. Every effort should be made to see that the manuscript itself contains no clues to the author's identity.

Manuscripts: Submit manuscripts in triplicate to the Editor, R. J. Herrnstein, *Psychological Bulletin,* Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, according to instructions provided below.

Instructions to Authors: Authors should follow the directions given in the *Publication Manual of the American Psychological Association* (2nd ed.). Instructions on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors are requested to refer to the "Guidelines for Nonsexist Language in APA Journals" (Publication Manual Change Sheet 2, *American Psychologist,* June 1977, pp. 487–494) before submitting manuscripts to this journal. All manuscripts should be submitted in triplicate and all copies should be clear, readable, and on paper of good quality. Dittoed copies are not acceptable and will not be considered. Authors are cautioned to carefully check the typing of the final copy and to retain a copy of the manuscript to guard against loss in the mail.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1979: $40 domestic, $42 foreign, $7 single issue. APA member rate: $15. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Dept., APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in January) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Printed in the U.S.A. Second-class postage paid at Arlington, Va., and at additional mailing offices.

Vol. 86, No. 4 July 1979

Psychological Bulletin

Paradoxical Tranquilizing and Emotion-Reducing Effects of Nicotine

David G. Gilbert
Florida State University

Investigations of the effects of nicotine on emotion and on indices of central and autonomic nervous system arousal are reviewed. Mechanisms that may account for the paradoxical finding that nicotine increases autonomic nervous system end-organ arousal yet has frequently been reported to reduce self-report and behavioral indices of emotion are evaluated. A number of mechanisms have been proposed, but none are backed by a convincing network of supportive data. The literature suggests that the mechanism(s) by which nicotine reduces indices of emotion is influenced by a wide variety of variables, including behavioral activity level, central nervous system arousal level, type of emotion, time since the administration of the drug, and the rate and dose of its administration. There is the need to demonstrate a paradigm that reliably causes nicotine-induced reductions of emotion. Once reliability has been established, further studies can vary the paradigm's potentially influential variables.

One of the most paradoxical and intriguing facts in the field of emotions is that smoking and nicotine cause significant increases in physiological arousal, especially of autonomic end organs, yet frequently reduce behavioral and self-report measures of emotion and cause people to report feelings of increased tranquility (Frith, 1971; Ikard, Green & Horn, 1969; Ikard & Tompkins, 1973). This paradox is underscored by the fact that a number of major popular theories of emotion (Lange & James, 1922; Mandler, 1975; Schachter & Singer, 1962) view increased autonomic arousal as an essential component of emotional processes. Possibly because of the relevance of this paradox to current theories of emotion and smoking, there have been an increasing number of studies that pertain to the resolution of these apparently contradictory findings.

Although some investigators have summarized various parts of the literature relevant to nicotine's emotion-decreasing and physiological - arousal - increasing properties, there has been no extensive review of studies selected for their relevance to the clarification of the physiological and psychological mechanisms that underlie these apparently contradictory effects. This review represents

---

An earlier version of this article was submitted to the Department of Psychology, Florida State University in partial fulfillment of the requirements for the PhD.

The author wishes to express his appreciation to Richard Hagen, Stephen West, Jack May, James Smith, and James Orcutt for their helpful comments.

Requests for reprints should be sent to David Gilbert, who is now at the North Central Florida Community Mental Health Center, 3615 S.W. 13th Street, Gainesville, Florida 32608.

an attempt to come closer to a determination of what these underlying mechanisms might be. First, studies demonstrating the paradox are reviewed. Then, mechanisms that have been proposed to explain the paradox are reviewed and evaluated in terms of their supporting evidence.

Paradox: Increased Autonomic Nervous System Activation With Decreased Emotion

Physiological Effects of Nicotine

The paradox exists only to the degree that nicotine produces simultaneous increases in physiological arousal and decreases in emotional experience or behavior. An examination of the effects of nicotine on physiological arousal and emotion is complicated by the lack of single definitive indices of either *emotion* or *physiological arousal,* concepts that are not homogeneous entities but instead are categories or constructs that represent a wide variety of different processes. Therefore, a variety of different indices of each of these categories are reviewed.

Effects on autonomic nervous system end organs. The smoking of one or two cigarettes or the administration of an equivalent amount of nicotine by other means typically causes significant sympathomimetic symptoms, the most notable of which include an increase in resting heart rate of from 5 to 40 beats per minute (Domino, 1973; Hill & Wynder, 1974; Roth, McDonald, & Sheard, 1945), increased blood pressure (5 to 20 mm of Hg), increased serum levels of epinephrine and adrenalcortical compounds (Herxheimer, Griffiths, Hamilton, & Wakefield, 1967; Hill & Wynder, 1974; Jarvik, 1970), and significant vasoconstriction (Herxheimer et al., 1967; Simon & Iglauer, 1967).

Almost all of the studies of the effects of nicotine on autonomic nervous system (ANS) arousal have maintained subjects in quiescent, nonemotional states prior to and after the administration of nicotine. Hence, there is some question about the extrapolation of these studies to subjects in active states, and evidence is quite meager in relation to this question. Two studies do support the notion that quiescent-state research is broadly applicable. These two studies found that high-nicotine cigarettes produced significantly greater increases in heart rate than low-nicotine cigarettes in moderately arousing emotional settings (Gilbert, 1978; Nesbitt, 1973). However, a study by Erwin (1971) found that heart rate did not increase after the onset of smoking in subjects who moved about, performing their daily routines.

Mechanisms by which nicotine induces these sympathetic symptoms have been reported to include direct effects of nicotine on the particular end organs and indirect effects brought about by nicotine-induced increases of serum epinephrine, which, in turn, stimulates these end organs (Jarvik, 1970; Meyers, Jawetz, & Goldfien, 1974).

Effects on central nervous system activity and structures. Contrary to the hypothesis that tobacco and nicotine produce central nervous system (CNS) tranquilizing effects, studies have typically shown that smoking-sized doses of nicotine (.005 to .1 mg per kg of body weight) produce electroencephalogram (EEG) activity characteristic of CNS arousal (alpha desynchronization, increased dominant alpha frequency, and reduction in total energy and variability of cortical EEG activity). The EEG arousal occurs immediately after the intravenous administration of nicotine and by the time the individual completes smoking a tobacco cigarette. Several studies using subhuman mammals have reported this initial phase of EEG arousal to be followed by a phase of EEG tranquilization that occurs about 5–15 minutes after the intravenous administration of the drug and lasts up to 30 minutes (Bhattacharya & Goldstein, 1970; Domino, 1967; Goldstein, Beck, & Mundschenk, 1967). However, no such secondary phase of EEG tranquilization has been shown to occur in humans. Philips (1971) reported that the smoking of a single cigarette resulted in a significant increase in EEG arousal (mean alpha power reduction) for up to 20 minutes. Similarly, Knott and Venables (1977) found that smoking two cigarettes resulted in EEG arousal (increased dominant alpha frequency) for the duration of their recording period (15 minutes). Nu-

merous other studies have also shown that smoking and intravenous and subcutaneous administration of nicotine induce EEG arousal in humans (Bickford, 1960; Hauser, Schwartz, Roth, & Bickford, 1958; Murphree, Pfeiffer, & Price, 1967; Wechsler, 1958) and in animals (Domino, 1973). However, one series of studies (Armitage, Hall, & Morrison, 1968) showed that the effects of small amounts of nicotine depended critically on dose and rate of administration when EEG and acetylcholine measures of cortical arousal were used in rats and cats. Furthermore, at some given doses and rates, it appeared that cortical arousal could be either increased or decreased, the effect varying from cat to cat.

On the other side of the coin, studies of smoking abstinence effects have consistently shown EEG sedation or depression effects following smoking deprivation (Itil, Ulett, Hsu, Klingenberg, & Ulett, 1971; Knott & Venables, 1977; Murphree & Schultz, 1968; Ulett & Itil, 1969). Deprivation also causes restlessness and dysphoria (Shiffman & Jarvik, 1976; Ulett & Itil, 1969). This simultaneous occurrence of EEG tranquilization and subjective dysphoria calls into question the validity of classical EEG measures as indices of emotional arousal. Furthermore, the occurrence of restlessness concurrently with EEG tranquilization calls into question the use of these EEG measures as indices of general arousal or activation level.

The EEG-activating effects of smoking have been interpreted in a number of studies (Bickford, 1960; Hauser et al., 1958; Wechsler, 1958) as resulting from the behavioral act of smoking rather than from the physiological effect of nicotine or other pharmacological substances present in tobacco. Studies systematically varying the nicotine content of the cigarette smoked and studies using intravenous administration of nicotine in humans are urgently needed. However, the animal studies mentioned above, which used smoking-sized intravenous doses of nicotine, have shown that nicotine induces classical EEG signs of arousal, a finding that strongly supports the view that nicotine is at least partially responsible for the EEG-arousing properties of smoking.

Less traditional, but possibly more sophisticated, measures of CNS arousal present a more complex picture of the effects of nicotine. Brown (1967) reported that intravenous nicotine produced a state of mixed arousal and sedation in cats. After nicotine administration, these cats demonstrated a behavioral state of drowsiness combined with a substained potential for alertness. Nicotine also produced an unusual partitioning of EEG activity. Consistent with arousal, there was a net decrease in energy content when all EEG frequencies were considered, but there was a sedative effect (increased energy content) among lower frequency ranges. Nicotine also led to a persistence of hippocampal theta frequency, which is characteristic of orienting responses. Finally, nicotine resulted in evoked responses in the hippocampus similar to those obtained during sleep. Vazquez and Toman (1967) also found that nicotine produced mixed and hard-to-explain effects on directly evoked brain potentials. In rabbits, they found that nicotine caused transient depression of the most prominent fast component but enhancement of the most prominent slow, negative component.

In studies with humans, the effects of cigarette smoking on cortical, average visually evoked potential and contingent negative variation indices of cortical arousal suggest that nicotine influences cortical processes in a complex manner that can be either arousing or sedating, depending on the nature of the stimulus and the personality characteristics of the individual. Smoking has produced an arousing (increased response amplitude) effect on average visually evoked potentials elicited by low-intensity stimuli and a sedative effect on potentials elicited by high-intensity stimuli (R. A. Hall, Rappaport, Hopkins, & Griffin, 1973). On the other hand, Ashton, Millman, Telford, and Thompson (1974) reported that smoking produced cortical arousal (increased contingent negative variation magnitude) in extraverts, but had the opposite effect (sedation) in introverts. This finding of the effects' dependence on personality is consistent with the findings of Brown (1967) and Armitage et al. (1968), which showed that

the effects of nicotine in cats and rats can either arouse or depress EEG, cortical evoked potentials, and cortical acetylcholine production, depending on the characteristic behavior of the individual animal.

In conclusion, nicotine typically causes increases in the more traditional and gross measures of general CNS arousal when the prenicotine arousal state of the organism is low. However, more detailed and sophisticated measures of arousal show that the effects of nicotine on CNS activity are multiple and depend on a variety of parameters. The picture is further complicated by reports of the onset of cortical sedation some 5 or 10 minutes after the onset of arousal in cats and rabbits, but not in humans, and by findings of differential effects of smoking and nicotine with different doses and behavioral predispositions. These conclusions indicate that future research should control for and systematically investigate the effects of these variables. Also, the findings of mixed arousal and sedation and of insignificant correlations between measures of CNS arousal indicate a need for simultaneous measurement of a variety of indices in different parts of the brain. Furthermore, this discordance of measures suggests that the concept of cortical or CNS arousal is inadequate to handle the data at hand. The inadequacy of the arousal construct is also attested to by the findings showing that smoking deprivation causes the simultaneous occurrence of EEG tranquilization and subjective restlessness–dysphoria.

These inadequacies and the fact that the effects of nicotine on the CNS are highly complex and dependent on numerous parameters suggest that a multiple interacting systems model is preferable to a unidimensional arousal model when one studies the effects of nicotine on brain activity and emotions. Such a systems approach suggests that consideration should be given to whether we are asking the right questions of the right parts of the CNS.

This systems approach also stresses the importance of the mutual influence of various bodily systems on each other. In this line, it should be noted that alterations in CNS activity brought about by nicotine have usually implicitly been assumed to be the direct result of nicotine's effects on CNS mechanisms and structures. This assumption may, however, be in error, since the primary sites of nicotine's action may result from at least two other sources. First, Bickford (1960) has reviewed a number of changes in peripheral bodily states (e.g., respiration rate, blood carbon dioxide and sugar levels, and drug-induced alterations of peripheral organ activity) that have been shown to influence EEG activity.

The important role that peripheral bodily and psychological states have in influencing EEG activity makes it clear that a major problem with studies of the effects of nicotine on CNS arousal is that nearly all the studies have maintained their subjects in a low-arousal, quiescent baseline state before the administration of nicotine. The absence of studies whose subjects were in emotional or other high-arousal conditions before the administration of the drug makes the relevance of these studies to the suggested tranquilizing properties of nicotine very questionable. Future studies that systematically vary prenicotine arousal level are urgently needed. A related difficulty is the problem of generalization from restricted laboratory states of low stimulus input and low activity level to more natural and emotional states.

Effects of nicotine on skeletal muscle activity. A number of findings are consistent with the possibility that some of nicotine's tranquilizing effects may result from the tendency of nicotine to reduce muscular tension and responses to emotional stimuli. A rapid, short-lived reduction in muscular tension in spastic patients was observed after each subject smoked a single cigarette (Webster, 1964). Also, a depression of the patellar tendon reflex (knee jerk) and associated musculature has been demonstrated in humans (Domino, 1973) and other mammals (see review by Silvette, Hoff, Larson, & Haag, 1962) following the administration of small amounts of nicotine. Furthermore, diminution of aversive-stimulation-induced jaw contractions and associated muscle potentials has been reported in human and

monkey subjects following smoking and the intravenous and oral administration of nicotine (Hutchinson & Emley, 1973). It seems reasonable to assume that reductions of muscular activity in response to emotional stimuli are subjectively experienced as emotion reducing or tension relieving; thus, these effects of nicotine on muscular activity are consistent with reports of nicotine's tranquilizing properties. Somewhat inconsistent with this hypothesis, the flexor reflex has been shown to be relatively unaffected by as well as enhanced by nicotine (Silvette et al., 1962). Nonetheless, overall, the evidence suggests that potentially significant reductions of various muscular activity or responses result when nicotine is administered.

Nicotine may reduce muscular activity and tension by causing CNS emotional centers to reduce emotional motoric output, or its action may be more peripheral, such as in the spinal cord, the neuromuscular junction, or the muscles themselves. Domino (1973) has cited evidence that suggests that the mechanisms of the reduction are complex and involve both central and peripheral components. Obviously, there is a need for further studies of the effects of nicotine on stress and emotion-induced muscular activity.

Summary. The physiological effects of smoking are essentially identical to those produced by the administration of equivalent doses of nicotine by other means. The predominant effect of nicotine on peripheral bodily structures is that of activation of end organs of the ANS in a sympathomimetic manner. It is true that the effects of nicotine on skeletal muscle action and on CNS activities are more complex and less well understood and established. However, the establishment of the paradox only requires that nicotine be shown to increase ANS arousal while simultaneously decreasing indices of emotion. The demonstration of the latter of these two requirements is now addressed.

Tranquilizing and Emotion-Reducing Effects of Nicotine

Self-reports. A majority of smokers report that they smoke to reduce negative affect or to achieve pleasurable relaxation (Ikard et al., 1969; Ikard & Tompkins, 1973). In a national sample of over 2,000 smokers, Ikard et al. found that 80% scored high on the factor of items indicating that they always or usually smoked for pleasurable relaxation. However, approximately one quarter of this large sample also reported smoking at times for the purpose of stimulation. Frith (1971) asked 59 men and 39 women to imagine themselves in 12 high-arousal situations and 10 low-arousal situations and to indicate on a 7-point scale what their desire for a cigarette would be. Smokers reporting smoking a large number of cigarettes per day indicated a high desire to smoke in both the high- and low-arousal situations. Less heavy smokers, however, could be broken into two groups, those who reported a great desire to smoke in low-arousal situations and those who reported a great desire to smoke in high-arousal situations. More men indicated a strong desire to smoke in the low-arousal circumstances, whereas more women reported a strong desire to smoke in the high-arousal situations.

Experimental studies have also shown that smoking and other means of administering nicotine reduce reports of experienced emotion. In the earliest of these studies, Johnston (1942) published a series of observations of the effects of nicotine on smokers and nonsmokers. Nonsmokers given 1.3-mg hypodermic injections of nicotine reported that an unpleasant light-headedness or muzziness appeared about 5 minutes after the injection. Smokers experienced the same sensations, but described them as pleasant. One patient receiving 4.3 mg of nicotine in water orally three times per day stated that it "steadied" her more than did phenobarbital. Heimstra (1973) described a series of better designed and controlled experiments in which groups of smokers were or were not allowed to smoke. If nicotine reduces emotions, withholding nicotine should increase emotions relative to not withholding nicotine. Smoking reduced the reported amount of mood change that occurred from before to after a series of tasks (6 hours of driving stimulation and 3 hours of pursuit rotary tracking, target detection, and reaction time tasks) and from

before to after a stressful movie. Smoking consistently reduced the mood factors of aggression and anxiety, as measured by the Mood Adjective Check List (Nowlis, 1965).

In a related series of studies, Frankenhaeuser and Myrsten and their associates have shown that in smoking-deprived (up to 15 or more hours) subjects, mental efficiency and subjective well-being are increased in a smoking situation as compared to a nonsmoking situation (Frankenhaeuser, Myrsten, & Post, 1970; Frankenhaeuser, Myrsten, Waszak, Neri, & Post, 1968; Myrsten, Post, Frankenhaeuser, & Johansson, 1972). In one of these studies (Myrsten et al., 1972), for example, smoking during reaction time tasks led to decreased self-estimates of irritation and boredom, as compared to the nonsmoking condition.

In an experiment by Ague (1973), subjects suspended smoking for 8 hours, rated their moods, and then smoked one of four cigarettes of varying nicotine content. An hour after smoking subjects again rated their moods. Subjects rated themselves as feeling significantly more pleasant and relaxed an hour after smoking high-nicotine cigarettes than they felt an hour after smoking low- or no-nicotine ones. Nicotine dose level did not result in statistically significant differences on other Mood Adjective Check List factors (Aggression, Anxiety, Surgency, Concentration, Fatigue, Vigor, and Sadness), but there was a nonsignificant tendency for nicotine to reduce anxiety and aggression.

In summary, it can be said that smoking has consistently led, in the published studies to date, to improved mood states in deprived smokers; but the question as to whether this improvement is caused by effects of nicotine or by the oral and manipulative activities involved in smoking has not been addressed by any of the studies other than that of Ague (1973). The probability that oral and manipulative activity is at least partially responsible for the calming effects of smoking seems very high in light of Freeman's (1948) work, which has shown that oral activity reduces anxiety measured by self-report. It is essential, therefore, that future studies control for the variety of smoking-associated behaviors by varying the nicotine content of the cigarettes smoked. The one reported study that did vary the nicotine content of cigarettes smoked (Ague, 1973) found that only two of the nine measured emotions (moods) were reduced or improved, and it is questionable whether these two self-report measures (*pleasantness* and *inner tension*) represented true or complete emotions. Even more desirable would be studies administering nicotine to humans by means other than smoking (e.g., intravenous, oral, or subcutaneous administration). Such means of manipulating nicotine level would control for taste differences among cigarettes of different nicotine contents and could more definitively demonstrate the emotion-reducing properties of the drug.

There is also a need for future studies that present emotional stimuli while varying nicotine level. Since most studies to date have not presented emotion-eliciting stimuli, they have in fact dealt with moods rather than with emotions. Finally, the question of whether nicotine also produces tranquilization in nonsmokers needs investigation.

Behavioral measures. Only two studies have investigated the effects of nicotine on behavioral measures of emotion in humans. The first of these investigations is that of Nesbitt (1973), who found that habitual smokers behaved less emotionally (were willing to endure stronger intensities of electrical shock) when smoking than when simulating smoking. He also found that smokers of a high-nicotine cigarette behaved less emotionally than smokers of a low-nicotine cigarette. He interpreted this study as suggesting that nicotine can reduce emotional behavior; however, the results are open to other plausible interpretations. First, Nesbitt noted that an alternative explanation is the possibility that smokers of the high-nicotine cigarette were satisfied with the cigarette they were allowed to smoke, whereas smokers in the no-cigarette condition were initially told they would be allowed to smoke but later were informed that they would not be allowed to smoke after all, so that the no-cigarette condition may have led to feelings of frustration and anger with an associated decrease in desire to please the ex-

perimenter by enduring the shocks. Similarly, low-nicotine cigarettes generally are considered to be less satisfying than cigarettes of normal nicotine content, and the former, therefore, may have produced feelings of frustration or dissatisfaction similar to the feelings of the no-nicotine subjects, although not as strong. A second alternative interpretation is related to the finding that nicotine increases detection thresholds for electrical shock (Mendenhall, 1925; Wenusch & Schöller, 1936). It may be, therefore, that nicotine decreases the perception of shock and thus accounts for the increased tolerance of shock. Thus, Nesbitt's results are consistent with the hypothesis that nicotine reduces emotional behavior, but they also are in line with alternative explanations.

In the second study, Schechter and Rand (1974) found that habitual smokers deprived of smoking were 37% more aggressive on a Buss (1961) aggression machine task than were subjects who were allowed to smoke. Unfortunately, the deprived smokers did not have the opportunity to engage in the same oral and manipulative behaviors as did smokers. As a result of this confound, no conclusion as to the role of nicotine can be reached.

In summary, too few studies have been reported and too many methodological problems are evident in the studies that are available to permit a definite statement about the role of nicotine in altering emotional behavior in humans. It can, however, be said that the studies reported to date suggest that nicotine may be able to reduce anxiety and aggression in habitual smokers, as determined by behavioral measures.

Emotion in animals. The leap from the discussion of emotions in man to emotions in subhuman animals is, of course, a questionable proposition, since there is no assurance that the "anger" or "anxiety" that a rat, for example, experiences corresponds in a significant manner to their counterparts in humans. Nonetheless, the effects of nicotine on behavioral measures of emotion in subhuman species have received increased attention during the last decade, and some strong trends appear to be developing in this area. First, animal studies have consistently shown that smoking-dose levels of nicotine reduce a variety of measures of aggression. For example, Silverman (1971) found a consistent reduction of aggression in both albino and hooded rats following small subcutaneous doses of nicotine. Sexual, investigatory, and submissive behaviors were also reduced very slightly, but general activity level was not significantly influenced. Consistent with these findings, Hutchinson and Emley (1973) reviewed a series of experiments that they and their co-workers have completed in recent years showing that the acute oral administration of small doses of nicotine (.04–.80 mg/kg) and the chronic oral administration of even smaller doses (as small as .002 mg/kg) reduced postshock biting in monkeys while simultaneously increasing preshock anticipatory motor behaviors. These two simultaneous effects were also produced in their studies by two tranquilizers (chlorpromazine and chlordiazepoxide) and have been said to be characteristic of tranquilizer-type compounds (Emley & Hutchinson, 1972). These researchers also found that withholding nicotine after chronic administration caused increases in their measures of aggression. The findings of less well-designed studies by Schechter (1974) and Kostowski (1968) are consistent with this general picture, in that they also showed nicotine-produced decrements of aggression in rats and ants, respectively.

Indices of fear and anxiety in animals have typically been reduced by nicotine, but these effects have been less consistent than the effects of nicotine on aggression. A number of researchers (Bovet, Bovet-Nitti, & Oliverio, 1966; Essman & Essman, 1971; Fleming & Broadhurst, 1975; Hutchinson & Emley, 1973; Morrison & Stephenson, 1972) have reported that nicotine reduces measures of anxiety (immobility, conditioned suppression, exploratory behavior) in a variety of animals. On the other hand, Davis, Kensler, and Dews (1973) and Driscoll and his coworkers (Driscoll, 1976; Driscoll & Bättig, 1970; Driscoll & Bättig, 1974) have reported nicotine-produced increases in indices of anxiety–fear (punishment-induced suppression of operant behavior and avoidance behavior).

Studies of the effects of nicotine on cor-

relates and indices of anxiety–fear in subhuman species are difficult to interpret not only because of apparently contradictory results but also because nicotine has a strong tendency to increase all forms of operant behavior, those motivated by a fear component and those that are not. Consistent with the emotion-reduction hypothesis, Hutchinson and Emley (1973) found what might be considered one of the best correlates of anxiety—conditioned suppression of positively reinforced behavior—to be reduced by low and intermediate doses of nicotine (.1–.4 mg/kg) administered subcutaneously in monkeys and rats. Also consistent with the suggestion that nicotine reduces anxiety are the findings of studies that have shown increased mobility and exploratory behavior in threatening environments, that is, in shock avoidance conditions (Fleming & Broadhurst, 1975; Morrison & Stephenson, 1972). These studies showing a reduction of the suppression of operant behavior, however, cannot be cited as strong support indicative of nicotine-induced reduction of anxiety–fear, since nicotine also increases a wide variety of operant behaviors not related to negative emotions.

The findings that have been interpreted as suggestive of nicotine-induced increases, rather than decreases, of fear and anxiety involve, for the most part, avoidance paradigms. Small and moderate doses (.5–.30 mg/kg) of nicotine, like caffeine and amphetamine, increase bar pressing in avoidance tasks (Balfour & Morrison, 1975; Davis et al., 1973). This increased rate of avoidance behavior is characteristically produced by stimulant drugs and does not nessarily indicate increased fear or anxiety, as evidenced by the fact that nicotine and stimulant drugs also increase operant behavior not related to negative emotions.

Studies testing the effects of nicotine on punishment-induced suppression of positively reinforced behavior in one case (Morrison, 1969) showed that nicotine and amphetamine reduced the punished behavior to below control values (which would be consistent with nicotine's having anxiety-increasing properties), whereas a tranquilizer (chlordiazepoxide) increased responding to nonpunished levels. However, a more recent study (Morrison & Stephenson, 1972) showed that both amphetamine and nicotine slightly increased behavior that had simultaneous rewarding and aversive properties. In this latter case, these two drugs appear to have an antianxiety effect.

In summary, a review of the effects of nicotine on behavioral measures of emotion in subhuman species suggests that the drug reduces a variety of aggressive behaviors, but its effects on other forms of emotional behavior are still open to question. In view of the mixed and hard-to-interpret effects of nicotine on behavioral indices of anxiety–fear, future research should investigate the importance of the variety of variables suggested earlier in this review in determining whether nicotine has stimulant or sedative properties. These variables include individual and strain differences, dose, paradigm type, rate of nicotine administration, intensity of emotion-inducing stimuli, preexperimental arousal level, and type of emotion.

Summary of the emotion-reducing effects of nicotine. The studies reviewed suggest that, at least in some situations, nicotine has emotion-reducing properties in man and animals and that nicotine deprivation in chronic users causes relative increases of emotion. All 12 of the reported studies with humans have been consistent with the hypothesis that nicotine has this effect. Furthermore, nicotine reduced indices of aggression in animals in all 7 studies available for review. And with the exception of avoidance measures, nicotine also has reduced correlates of anxiety–fear in animals (even though these studies are open to alternative explanations).

There is moderately strong evidence for the existence of the paradox: Nicotine significantly increases arousal of essentially all aspects of the ANS, frequently increases indices of CNS arousal, and at the same time frequently results in reduced negative emotional feelings and behavior and in increased feelings of tranquility and pleasure. The question as to whether nicotine has tranquilizing properties in humans who are not chronic users of the drug has not been answered. The studies involving human subjects have a number of methodological flaws,

but are consistent with the animal literature in suggesting that nicotine reduces indices of emotion.

Mechanisms Proposed to Explain the Paradox

It is important to note that the different mechanisms that have been proposed as explanations of nicotine's probable emotion-reducing properties are not necessarily mutually exclusive. Most of the mechanisms deal with only one level of analysis, so that different mechanisms may account for only part of the picture in a manner similar to the "blind men and the elephant" analogy. In addition, it should be kept in mind that since nicotine produces a large number of direct and indirect physiological changes, the different proposed mechanisms concern themselves with the primary sites of nicotine's action, this primary action then leads to immensely complex chains of further action involving hormonal, neural, behavioral, and experiential changes.

Addiction, withdrawal, and related processes cannot be considered adequate explanations of the paradox without specifically noting by what mechanism nicotine decreases emotion while simultaneously increasing arousal. Inconsistent with the view that the paradox is a manifestation of an addictive process, each dose of nicotine produces transient increases of arousal relative to chronic arousal levels (Ague, 1973; Gilbert, 1978). Also, the paradox has been shown to appear in individuals with no prior experience with the drug (Hutchinson & Emley, 1973). However, the opponent-process theory of addictions and emotions (Solomon & Corbit, 1973) is helpful in that from it one can infer that nicotine may reduce emotion by one mechanism after its chronic administration and by another mechanism in individuals without extensive prior use. Furthermore, the opponent-process theory of addictions leads one to speculate that individuals may adapt to the heightened arousal levels brought about by nicotine and that depriving the addict results in a decrease of arousal that may be subjectively experienced as aversive and thus cause the subject to respond with more emotion. Consistent with this hypothesis, Heimstra (1973) has reported data showing that deprived smokers report simultaneous decreases in arousal (increased fatigue) and increases in aggression and anxiety.

This review is limited to the effects of nicotine. Therefore, it does not consider oral, manipulative, attentional, attributional, and other psychological and behavioral factors associated with smoking independent of nicotine, even though these factors may be very important in producing the satisfying effects of smoking.

CNS Mechanisms

Neurophysiological model. Miller (1973) pointed to a finding in his laboratory that showed that chemomicrostimulation-induced activation of muscarinic neurons in certain portions of the cat brain elicited aggression. He suggested that if the effects of muscarinic and nicotinic central cholinergic neural systems are antagonistic like those of alpha and beta adrenergic systems, then nicotine's calming effect would be explained, since nicotine stimulates the nicotinic neural system and this system tends to inhibit the aggression-producing muscarinic system. Consistent with this hypothesis, Silverman (1971) has noted that the behavioral effects of nicotine are similar to those of benactyzine, which antagonizes the muscarinic effects of acetylcholine. However, this study provides only indirect evidence, and the literature does not appear to provide more direct tests of the proposed nicotinic–muscarinic antagonism. Thus, the status of the hypothesis must be considered speculative and in need of systematic study.

This proposed inhibition of muscarinic aggression circuits by nicotine may prove to account for nicotine's ability to reduce certain forms of aggression, but it does not account for the drug's apparent ability to lessen emotions other than aggression. It is not clear whether nicotine reduces different emotions by altering emotion-specific neurophysiological pathways or by influencing a pathway or process common to all emotions. If the former were true, the effects of nicotine on different emotions would be expected to

vary more widely from emotion to emotion than if the latter were true.

A review of the literature (Avis, 1974) makes it abundantly evident that the role of neurotransmitter systems in the elicitation and inhibition of emotion is extremely complex. This regulation of emotions is a function of relationships between numerous transmitter systems rather than endogenous activity in one isolated system. This and other reviews also demonstrate that different emotions and different forms of the same emotion are mediated in the CNS by partially independent neurophysiological systems (Avis, 1974; Gittelman-Klein, & Klein, 1971; Heiser & DeFrancisco, 1976).

The suggestion that nicotine inhibits muscarinic aggression circuits is consistent with Silverman's (1971) findings showing nicotine's main effect on social behavior in rats to be a reduction of aggression. It also conforms to the fact that in the nicotine studies to date, aggression is the one emotion that has most consistently been reduced. This model does not, however, explain why it is that many smokers report feeling tranquilized by smoking in spite of smoking's ANS-arousal-producing effects.

It can be inferred from Miller's (1973) neurophysiological model that nicotine may reduce both behavioral and subjective indices of emotion by a direct neurophysiological inhibition of the impulse to behave in an emotional manner. It has been suggested (Arnold, 1960) that the impulse to respond to an emotion-producing stimulus is of greater importance in the subjective experience of emotion than is the perception of ANS arousal.

Cortical sedation model. Recently, Eysenck (1973) proposed that the effects of nicotine depend on the degree of arousal in the cerebral cortex: When arousal is high, the effects on the cortex are depressant; when arousal is low, the effects are stimulating. During the highly arousing states that accompany most emotions, cortical arousal is high (Lindsley, 1970); hence, the effects of nicotine on the cortex during emotional states are depressant (i.e., tranquilizing). Decreased cortical arousal is assumed to cause decreased emotion. This model can account for nicotine's tranquilizing properties if one assumes that CNS sedation results in decreased CNS response to and perception of ANS-arousal-related input.

The strongest support Eysenck offered for his hypothesis that nicotine has bidirectional effects on arousal came from Armitage and his co-workers (Armitage et al., 1968; Armitage, Hall, & Sellers, 1969). These researchers found that the effects of nicotine administered to cats and rats at a frequency and dosage per body weight corresponding to human smoking resulted in cortical (EEG and acetylcholine level) and behavioral arousal (lever-pressing rate) in some animals of each species and in tranquilization in others.

The studies mentioned earlier in this review that showed that the effects of smoking and nicotine on cortical activity can be either arousing or depressing provide indirect evidence for Eysenck's hypothesis. Also consistent with this bidirectional view is the finding that nicotine tends to increase sensory detection thresholds when CNS arousal is high, but decreases them when CNS arousal is low (Mendenhall, 1925).

Finally, a behavioral study of Ashton and Watson (1970) offers support for Eysenck's bidirectional model. These researchers found that smokers smoked more when underaroused (resting) and when highly aroused (high-stress situation) than they did when only slightly aroused, which suggests that underaroused subjects smoked for stimulation and overaroused subjects for tranquilization.

Eysenck did not speculate as to what physiological mechanism might underlie the suggested bidirectional effects of nicotine on cortical arousal. He did, however, note that this hypothesis is consistent with the conclusion of Rachman (1969) that arousal is an inverted-U-shaped function of the arousal-producing conditions (Eysenck, 1973, p. 136).

There is very little research that relates directly to Eysenck's hypothesis. A test of this proposal would be to expose subjects to an emotion-producing situation and, after cortical arousal was high, to administer nico-

tine while continuing to monitor CNS indices of arousal.

Interacting CNS mechanisms. A possibility not discussed by Eysenck is that there are two brain arousal systems that interact, so that arousal in one system tends to inhibit arousal in the other system. Routtenberg (1968) has made a strong argument for the likelihood of the existence of two such arousal systems (limbic and reticular activating) and has noted evidence that arousal of either one of these systems leads to cortical arousal but that the simultaneous arousal of both of them leads to cortical tranquility. Consistent with this hypothesis, Barnes (1966) showed that either of two stimulants (eserine or amphetamine) caused EEG arousal (fast waves), but surprisingly, the administration of both reinstated tranquilization (slow wave EEG activity). If emotional arousal is based primarily on the activation of one of these two systems, the tranquilizing properties of nicotine may be a result of the inhibition of this system by nicotine's arousing of the other system. A replication of Barnes's study with nicotine, as well as eserine and amphetamine, would be a test of this hypothesis.

Related to the idea of mutually inhibitory arousal systems are the findings of Friedman, Horvath, and Meares (1974), which showed that tobacco (but not nontobacco) cigarette smokihg significantly increased the rate of habituation of EEG alpha desynchronization to a series of 90-dB sound pulses. These researchers proposed that since theoretical models suggests that habituation depends on active inhibitory and excitatory mechanisms that oppose each other (Sokolov, 1960), it is possible that nicotine causes a dislocation of the usual relationship, so that inhibitory mechanisms are more fully activated without a simultaneous reduction in CNS excitation. This model is consistent with the earlier discussed findings that showed that nicotine increases some measures of cortical arousal while reducing others. It is also consistent with the EEG and self-report data, reviewed earlier, that suggest that nicotine increases the state of mental alertness while simultaneously producing a state of emotional tranquility. This model suggests that nicotine causes an individual first to become aware of, but soon to pay little attention to, emotion-related stimuli, thoughts, and ANS end-organ activity.

Studies that simultaneously monitor CNS indices of arousal and responses to external emotional input would be relevant to this hypothesis. A modified replication of the study by Friedman et al., in which nicotine would be administered intravenously or by other nonsmoking means, would provide a significant amount of credibility to this hypothesis.

Another model has been suggested by Goldstein et al. (1967). It is based on their studies of the effects of nicotine on simultaneous quantitative measures of cortical electrical activity in the rabbit brain. They interpreted their results as indicating that nicotine causes an impressive diminution of mutual involvement (EEG covariance) between the cortex and the hippocampus, and between the cortex and the reticular formation. These findings, they suggested, may imply a decrease in the activity of inhibitory mechanisms operating between cortical and subcortical sites. Thus, for instance, nicotine may relax excessively rigid control by the neocortex of subcortical structures (and vice versa). This model can explain the tranquilizing effects of nicotine if one assumes that either subcortical input to cortical structures or phasic deviation from the ANS-arousal baseline, rather than the baseline itself, determines the intensity of emotional responses. The nicotine from a single cigarette sustains high levels of ANS end-organ arousal for up to 30 minutes. On the other hand, startle responses and many other emotion-related responses that are likely to be mediated by interactions between cortical and subcortical structures usually operate in a much smaller time frame.

Pleasure-center-stimulation models. Eysenck (1973) has suggested an extension of his proposed cortical sedation explanation of nicotine's paradoxical effects. The additional proposal is based on the suggested effects of nicotine on pleasure and aversion systems in the brain. First, granting the very speculative nature of the hypothesis, Eysenck started by reminding the reader that three

different hedonic systems have been described (Berlyne, 1971; Olds & Milner, 1954; Olds & Olds, 1965). The first two are the primary reward and aversion systems; the stimulation of either of these leads to familiar signs of increased arousal, including increased heart rate, high-frequency EEG waves, and increased bodily movement; the third is the secondary reward system, the activation of which results in de-arousal. It is suggested that the secondary reward system produces rewarding effects indirectly, that is, by inhibiting the aversion system, which in turn releases the primary reward system from inhibition by the aversion system. Eysenck (1973, p. 140) proposed that nicotine administered in emotionally arousing situations may activate the secondary reward system and through it deactivate the aversion system. On the other hand, he also suggested that in low-arousal situations nicotine may activate the primary reward system directly. Eysenck offered no direct support for this model, but did note that it is open to experimental falsification.

In contrast with Eysenck's hypothesis, the model proposed by Jarvik (1970, 1973) assumes that nicotine's emotion-reducing properties are a product of its presumed ability to stimulate primary rather than secondary reward centers. Nicotine, like amphetamine, is assumed to release stores of norepinephrine and other catecholamines in primary reward areas of the brain such as the medial forebrain bundle, the results being an improvement in mood and a reinforcement of behavior associated with the administration of the drug. This interpretation is consistent with the catecholamine theory of emotion (Schildkraut & Kety, 1967), which assumes that catecholamine activity in the brain, particularly norepinephrine activity but possibly also dopamine and epinephrine activity, is responsible for sustaining positive affect. The pleasure-center-stimulation model can be inferred to assume that a hedonic process is an important component of emotion and that nicotine's purported pleasure-increasing properties more than compensate for its ANS-arousing properties.

Jarvik's and, to a lesser extent, Eysenck's hypotheses are not supported by the finding (Schuster, 1970) that amphetamine increased rather than decreased cigarette-smoking frequency in habitual smokers. In his discussion of these findings, Schuster pointed out that it is difficult to understand why smoking increased, since one assumes that there is at least some satiation of the central reward mechanisms by amphetamine, leading to a diminished need for nicotine intake.

On the other hand, support for the pleasure-center-stimulation model is given by findings that suggest that nicotine releases acetylcholine in the brain (Armitage, 1973; Essman, 1973; Knapp & Domino, 1962). Finally, Armitage (1973) has said that the findings of a recent study by G. H. Hall and Turner (1972) suggest that nicotine does in fact release noradrenaline in the specific areas of diencephalic pleasure centers, thus providing an important link to Jarvik's (1970, 1973) hypothesis that nicotine produces its effects by stimulating pleasure centers in the brain and causing or facilitating the release of catecholamines in these centers. With cats, intravenously injected nicotine (.002 mg/kg every 30 sec) caused an increased release of ^{3}H-noradrenaline into the effluent of the third cerebral ventrical, as did the administration of cigarette smoke directly into the lungs.

Also consistent with the pleasure-center-stimulation model, but inconsistent with arousal formulations of emotion, is that amphetamine, a CNS and ANS stimulant that also appears to stimulate pleasure centers (Jarvik, 1970, p. 188; Ray, 1972, p. 167), causes subjective assessments of relaxation that are far in excess of responses relating to nervousness (Martin, Sloan, Sapira, & Jasinski, 1971). Furthermore, it is a well-accepted fact that amphetamine, methylphenidate, and other CNS stimulants have paradoxical tranquilizing effects on hyperactive children (Satterfield, Cantwell, & Satterfield, 1974) and on some adults who manifest overreactivity and other symptoms of the hyperactive syndrome (Wood, Reimherr, Wender, & Johnson, 1976). Finally, small doses of amphetamine have been shown to frequently reduce aggression in laboratory animals, whereas large doses have been

shown to increase it (for a review, see Allen, Safer, & Covi, 1975).

In summary, several findings are consistent with the hypothesis that nicotine's emotion-reducing properties are at least partially a function of its suggested ability to stimulate CNS pleasure centers. The findings of G. H. Hall and Turner (1972) offer some direct evidence suggesting that nicotine stimulates pleasure center activity. However, the question of whether this increment in activity increases pleasure, whether nicotine increases such activity during emotional states, and whether such activity is related causally or only correlationally to reduced emotions have not been investigated. Amphetamine appears to stimulate pleasure centers and, like nicotine, reduces several indices of anxiety and aggression, which suggests that there may be a causal relation between pleasure and the inhibition of emotions. However, a great deal of further investigation is needed before an accurate assessment of the pleasure-center-stimulation model can be made.

Mechanisms Based on Altered Perceptions of Peripheral Activity

One of the three mechanisms based on altered perceptions of bodily activity discussed below may account for nicotine's reduction of emotions in spite of increased ANS activity. The first of these mechanisms is based on the findings of Wenusch and Schöller (1936) and Mendenhall (1925), which showed an increase in the detection threshold for electrical shock following the smoking of a regular tobacco cigarette but not following the smoking of a nicotine-free cigarette. These results suggest that such smoking-induced increases of sensory thresholds may decrease subjects' awareness of their autonomic arousal in spite of an actual increase in ANS activity. If emotion is more a function of consciously perceived than of actual autonomic arousal, then the paradox is resolved. This model makes no assumptions as to whether the elevation in threshold is caused by a general cortical sedation similar to that suggested by Eysenck (1973) or whether it is a result of some more specific process such as that suggested by the neurophysiological model discussed earlier in this review.

The second of the proposed mechanisms based on altered perceptions of bodily activity assumes that nicotine's tranquilizing properties are a consequence of muscular-action-reducing properties of nicotine. Research shows that while nicotine increases ANS arousal, it reduces reflexive muscular activity (patellar and startle responses) in humans and primates (Domino, 1973; Hutchinson & Emley, 1973) and reduces resting muscle tone levels in spastic patients (Webster, 1964). Hence, it may be that people who report that smoking tranquilizes them attend to muscular changes, the tranquilizing effects of which overshadow the effects of smoking-induced increases of autonomic arousal. It appears that the subjective experience of emotion is a positive function of facial and skeletal muscular activity as well as of ANS arousal. If this hypothesis is true, nicotine can be expected to reduce emotions with a large muscular component more than those with a relatively greater autonomic arousal component.

The final of the three proposed mechanisms was suggested by Schachter (1973, p. 152) in a direct attempt on his part to explain the paradoxical tranquilizing effects of smoking. This mechanism is based on the *law of initial values*, the finding that "the magnitude of response is related to prestimulation level" in such a manner that "high autonomic excitation preceding stimulation is correlated with low autonomic reactivity upon stimulation" (Lacey, 1956, p. 156). The argument is that since nicotine leads to arousal, the additional arousal induced by an emotional situation is less with nicotine than without it. If the intensity of an emotion is a positive function of the deviation of autonomic activity from its baseline, this explanation is plausible. At the present time, however, there appear to be no data either strongly in support of or contrary to this ANS deviation model of emotion.

Altered Cue Value of Peripheral (Bodily) Activity

It may be possible to feel physiologically aroused and yet feel unemotional and tranquil. Schachter (1973), in the same article in which he proposed the law-of-initial-values solution to the paradox, noted the possibility of an alternative model. This alternative assumes that habitual smokers experience their body as reacting with the same sympathetic symptoms when they smoke as it reacts when they experience an emotion; therefore, to the degree that they attribute their bodily arousal to smoking, they should be less emotional; that is, in an emotional situation a smoker feels emotion only to the degree that arousal (both emotional stimulus induced and nicotine induced) is attributed by the person to an emotional stimulus. To the degree that arousal is attributed to smoking (or other nonemotional sources), emotion is not felt (i.e., there is tranquilization). Unfortunately, this proposal by Schachter seems unlikely to be a general solution to the paradox, since subhuman species with little or no experience with nicotine show reductions of emotional behavior and it appears that there is no reason for such animals to attribute their arousal to a source other than the emotion-producing stimulus.

Impulse Reduction Model

It may be that a person can feel autonomic arousal and yet feel relaxed, tranquil, and unmotivated. This model does not assume, as does Schachter's, that misattribution of arousal to a nonemotional source is the critical step; nor does the model require prior experience with nicotine. Instead, it suggests that nicotine may reduce the motivation, the impulse-to-move component of the emotional complex, possibly in a manner similar to the unexplained mechanism by which morphine reduces pain. Ray (1972) has noted that many laboratory reports suggest that morphine has no effect on pain threshold and that it does not impair conduction in peripheral nerves. Following the administration of an analgesic dose of morphine, some patients report that they still notice pain but that it is no longer aversive; nor does the pain demand attention or an impulse to action (Ray, 1972). Similarly, nicotine may not alter or may actually increase the subjective perception of ANS arousal, yet may decrease the attention an individual gives these symptoms and may reduce the impulse to respond to this bodily activity and to external emotion-producing stimuli.

A reduction of the impulse to respond to the emotion-producing stimulus would result in lessened behavioral measures of emotion and, according to Arnold's (1960) well-known theory of emotion, would reduce subjectively experienced emotion as well, since Arnold argued that emotion is best conceived of as the felt urge toward or away from an object perceived as good or bad. Furthermore, studies showing that nicotine reduces emotional muscular reflexes in man and monkeys (Domino, 1973; Hutchinson & Emley, 1973) are consistent with this impulse reduction model, as is Miller's (1973) neurophysiological hypothesis of antagonistic nicotinic and muscarinic neural systems, which was discussed earlier. Finally, the recent discovery of naturally occurring opiatelike substances (endorphins and enkephalins) in the brain and their reported emotion-reducing effects (Arehart-Treichel, 1978) adds to the plausibility of this mechanism.

Glucocorticoid–ACTH Model

It is possible that nicotine's emotion-reducing effects are mediated by the nicotine-induced release of glucocorticoids from the cortex of the adrenal gland. Glucocorticoids have been reported to reduce a variety of indices of emotion (Di Giusto, Cairncross, & King, 1971; Endroczi, Lissak, Fekete, & DeWied, 1970; Levine, 1971), and nicotine, in small, smoking-sized doses, has been shown to cause increased serum levels of glucocorticoids (Hill & Wynder, 1974; Kershbaum, Pappajohn, & Bellet, 1968). These two facts suggest that nicotine-induced reductions of emotional reactions are mediated and caused by nicotine-induced increases in serum glucocorticoids. Furthermore, nicotine and glucocorticoids have parallel effects on the following indices of

emotion, sensory detection threshold, and higher cognitive functions:

1. They reduce measures of emotion (Di Giusto et al., 1971; Levine, 1971; Schechter & Rand, 1974; Silverman, 1971).
2. They reduce startle responses (Hutchinson & Emley, 1973; Levine, 1971).
3. They facilitate habituation of the EEG alpha-desynchronization orienting response (Endroczi et al., 1970; Friedman et al., 1974).
4. They increase sensory detection thresholds (Henkin, 1970; Mendenhall, 1925).
5. They increase higher integrative functioning such as memory, learning, discrimination, and accuracy (Andersson & Post, 1974; Geller, Hartmann, & Blum, 1971; Henkin, 1970; Levine, 1971; Myrsten et al., 1972; Nelsen & Goldstein, 1972).

Indirect support for the glucocorticoid model is provided by evidence suggesting that nicotine's emotion-reducing and cognition-facilitating effects occur maximally 15 to 45 minutes after the administration of nicotine (Andersson, 1975; Andersson & Post, 1974), at the same time at which nicotine-induced glucocorticoid blood levels are highest (Hill & Wynder, 1974).

It should be noted that glucocorticoid levels in the blood interact in a complex manner with ACTH levels and sex hormones and that these hormones have been shown to influence indices of emotion and learning (Di Giusto et al., 1971; Leshner et al., 1973; Levine, 1971). Since ACTH from the pituitary gland stimulates secretion of glucocorticoids from the adrenal cortex, it seems likely that nicotine increases glucocorticoid levels by increasing blood levels of ACTH. If nicotine in fact does increase blood levels of ACTH, it would be of significance, since ACTH has been shown to reduce aggressiveness in mice (Leshner et al., 1973). Therefore, it may be more appropriate to consider a general pituitary, adrenal model of the emotion-reducing effects of nicotine.

Summary and Conclusions

A number of contemporary theories of emotion assume that emotions are largely a positive function of ANS arousal. However, a substantial number of studies have shown that nicotine increases heart rate, blood pressure, and numerous other indices of autonomic arousal; yet rather than producing expected increases of emotional behavior and feelings, it usually decreases emotions. A more detailed analysis reveals that the effects of nicotine on physiology and emotions are quite complex. First, nicotine typically increases ANS arousal, but the findings of one investigation suggest that this increase may depend on predrug arousal level. Likewise, nicotine's effects on classical CNS measures of arousal have typically been in the direction of increased arousal, while its effects on more specific and sophisticated CNS indices have been mixed and have depended on such factors as individual differences, rate of nicotine administration, and locus of brain activity monitored. One must question the relevance of the studies to date on the effects of nicotine on CNS activity, since they have not used emotionally aroused subjects, but have maintained subjects in pre- and postnicotine states of low arousal and inactivity. There is an obvious need for a systematic series of studies that monitor a variety of CNS, ANS, and muscular indices while manipulating and controlling personality, time, arousal, emotional, and drug parameters.

In spite of the fact that several indices of emotion have consistently been reduced by the administration of nicotine, the question as to whether all types and intensities of emotion are reduced has not been answered. It seems likely that nicotine's emotion-reducing effects vary from emotion to emotion, across various indices of the same emotion, and with a variety of other factors. The systematic investigation of the effects of nicotine on the whole spectrum of emotional dimensions could add significantly to the understanding not only of nicotine's paradoxical effects but also of emotional processes in general.

A number of mechanisms that may account for nicotine's paradoxical tranquilizing effects have been proposed, but none are backed by a convincing network of supportive data. It seems evident that aside from

a serendipitous major breakthrough, a great number of studies testing a large variety of proposed theories will be needed before the mechanism or mechanisms underlying the paradox are determined. Different proposed mechanisms are not necessarily mutually exclusive. The impulse reduction model, for example, is an attempt to integrate several low-level mechanisms into a higher order or more complete picture of the emotion-reducing properties of nicotine. This model is consistent with muscular-tension-reduction, pleasure center, and neurophysiological models, but adds a higher level of conceptual integration that allows one to see that the research on nicotine and emotions is consistent with Arnold's (1960) impulse theory of emotions, whereas it is inconsistent with many other conceptions of emotion.

The literature surveyed in the present review suggests that the mechanism(s) by which nicotine reduces emotion is influenced by a wide variety of variables, including behavioral activity level, CNS arousal level, personality, type of emotion, time since administration of drug, and rate and dose of nicotine administration. Any mechanism that is suggested to fully explain nicotine's tranquilizing properties must account for the influence of these variables; but unfortunately, of the mechanisms reviewed earlier, only Eysenck's (1973) proposal attempts to deal with their contribution. The most promising approach for future research would seem to be for studies to systematically manipulate and control these influential variables in evaluations of a spectrum of suspected lower and higher order mechanisms. A major first step would be the development of a paradigm with specific parameters that reliably demonstrates the nicotine-induced reductions of emotion. Once this paradigm has been established further studies can vary its potentially influential variables, thus contributing to a clearer understanding of the drug's paradoxical effects.

References

Ague, C. Nicotine and smoking: Effects upon subjective changes in mood. *Psychopharmacologia,* 1973, *30,* 323–328.

Allen, R. P., Safer, D., & Covi, L. Effects of psychostimulants on aggression. *Journal of Nervous and Mental Disease,* 1975, *160,* 138–145.

Andersson, K. Effects of cigarette smoking on learning and retention. *Psychopharmacologia,* 1975, *41,* 1–5.

Andersson, K., & Post, B. Effects of cigarette smoking on verbal rote learning and physiological arousal. *Scandinavian Journal of Psychology,* 1974, *15,* 263–267.

Arehart-Treichel, J. The pituitary's powerful protein. *Science News,* 1978, *114,* 374–381.

Armitage, A. K. Some recent observations relating to the absorption of nicotine from tobacco smoke. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives.* Washington, D.C.: V. H. Winston, 1973.

Armitage, A. K., Hall, G. H., & Morrison, C. F. Pharmacological basis for the tobacco smoking habit. *Nature,* 1968, *217,* 331–334.

Armitage, A. K., Hall, G. H., & Sellers, C. M. Effects of nicotine on electrocortical activity and acetylcholine release from the rat cerebral cortex. *British Journal of Pharmacology,* 1969, *35,* 157–160.

Arnold, M. G. *Emotion and personality.* New York: Columbia University Press, 1960.

Ashton, H., Millman, J. E., Telford, R., & Thompson, J. W. The effect of caffeine, nitrazepam, and cigarette smoking on the contingent negative variation in man. *Electroencephalography and Clinical Neurophysiology,* 1974, *37,* 59–71.

Ashton, H., & Watson, D. W. Puffing frequency and nicotine intake in cigarette smokers. *British Medical Journal,* 1970, *3,* 379–381.

Avis, H. H. The neuropharmacology of aggression: A critical review. *Psychological Bulletin,* 1974, *81,* 47–63.

Balfour, D. J., & Morrison, C. F. A possible role for the pituitary-adrenal system in the effects of nicotine on avoidance behavior. *Pharmacology, Biochemistry, and Behavior,* 1975, *3,* 349–354.

Barnes, C. D. The interaction of amphetamine and eserine on the EEG. *Life Sciences,* 1966, *5,* 1897–1902.

Berlyne, D. E. *Aesthetics and psychobiology.* New York: Appleton-Century-Crofts, 1971.

Bhattacharya, I. C., & Goldstein, L. Influence of acute and chronic nicotine administration on intra- and inter-structural relationships of the electrical activity in the rabbit brain. *Neuropharmacology,* 1970, *9,* 109–118.

Bickford, R. Physiology and drug action: An electroencephalographic analysis. *Federation Proceedings,* 1960, *19,* 619–625.

Bovet, D., Bovet-Nitti, F., & Oliverio, A. Effects of nicotine on avoidance conditioning of inbred strains of mice. *Psychopharmacologia,* 1966, *10,* 1–5.

Brown, B. B. Relationship between evoked response changes and behavior following small

doses of nicotine. *Annals of the New York Academy of Sciences*, 1967, *142*, 190–200.

Buss, A. H. *The psychology of aggression*. New York: Wiley, 1961.

Davis, T. R. A., Kensler, C. J., & Dews, P. B. Comparison of behavioral effects of nicotine, d-amphetamine, caffeine and dimethylheptyl tetrahydrocannabinol in squirral monkeys. *Psychopharmacologia*, 1973, *32*, 51–65.

Di Giusto, E. L., Cairncross, K., & King, M. G Hormonal influences on fear-motivated responses. *Psychological Bulletin*, 1971, *75*, 432–444.

Domino, E. G. Electroencephalographic and behavioral arousal effects of small doses of nicotine: A neuropsychopharmacological study. *Annals of the New York Academy of Sciences*, 1967, *142*, 216–244.

Domino, E. G. Neuropsychopharmacology. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives*. Washington, D.C.: V. H. Winston, 1973.

Driscoll, P. Nicotine-like behavioral effect after small dose mecamylamine in Roman high-avoidance rats. *Psychopharmacologia*, 1976, *46*, 119–121.

Driscoll, P., & Bättig, K. The effect of nicotine and total alkaloids extracted from cigarette smoke on avoidance behavior in rats under extinction procedure. *Psychopharmacologia*, 1970, *18*, 305–313.

Driscoll, P., & Bättig, K. Effects of nicotine on the shuttlebox behavior of trained guinea pigs. *Psychopharmacologia*, 1974, *38*, 47–54.

Emley, G. S., & Hutchinson, R. R. Basis of behavioral influence of chlorpromazine. *Life Sciences*, 1972, *11*, 43–47.

Endroczi, E., Lissak, T., Fekete, T., & DeWied, D. Effects of ACTH on EEG habituation in human subjects. *Brain Research*, 1970, *32*, 254–261.

Erwin, C. W. Cardiac rate responses to cigarette smoking: A study using radiotelemetry. *Psychophysiology*, 1971, *8*, 75–81.

Essman, W. B. Nicotine-related neurochemical changes: Some implications for motivational mechanisms and differences. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives*. Washington, D.C.: V. H. Winston, 1973.

Essman, W. B., & Essman, S. G. Cholinergic mechanisms and avoidance behavior acquisition: Effects of nicotine in mice. *Psychological Reports*, 1971, *29*, 987–993.

Eysenck, H. J. Personality and the maintenance of the smoking habit. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives*. Washington, D.C.: V. H. Winston, 1973.

Fleming, J. C., & Broadhurst, P. L. The effects of nicotine on two-way avoidance conditioning in bidirectionally selected strains of rats. *Psychopharmacologia*, 1975, *42*, 147–152.

Frankenhaeuser, M., Myrsten, A. L., & Post, B. Psychophysiological reactions to cigarette smoking. *Scandinavian Journal of Psychology*, 1970, *11*, 237–245.

Frankenhaeuser, M., Myrsten, A. L., Waszak, M., Neri, A., & Post, B. Dosage and time effects of cigarette smoking. *Psychophа rmocologia*, 1968, *13*, 311–318.

Freeman, G. L. *The energetics of human behavior*. Ithaca, N.Y.: Cornell University Press, 1948.

Friedman, J., Horvath, T., & Meares, R. Tobacco smoking and a "stimulus barrier." *Nature*, 1974, *248*, 455–456.

Frith, C. D. Smoking behavior and its relation to the smoker's immediate experience. *British Journal of Social and Clinical Psychology*, 1971, *10*, 73–78.

Geller, I., Hartmann, R., & Blum, K. The effects of nicotine, nicotine monomethiodide, lobeline, chlordiazepoxide, meprobamate and caffeine on a discrimination task in laboratory rats. *Psychopharmacologia*, 1971, *20*, 355–365.

Gilbert, D. G. Extraversion, type of smoker and the effects of nicotine on physiological and self-report measures of emotion (Doctoral dissertation, Florida State University, 1978). *Dissertation Abstracts International*, 1978, *39*, 1477B–1478B. (University Microfilms No. 78-15,456)

Gittelman-Klein, R., & Klein, D. F. Controlled imipramine treatment of school phobia. *Archives of General Psychiatry*, 1971, *25*, 204–207.

Goldstein, L., Beck, R. A., & Mundschenk, D. L. Effects of nicotine upon cortical and subcortical electrical activity of the rabbit brain: Quantitative analysis. *Annals of the New York Academy of Sciences*, 1967, *142*, 170–180.

Hall, G. H., & Turner, D. M. Effects of nicotine on the release of ^{3}H-noradrenaline from the hypothalamus. *Biochemical Pharmacology*, 1972, *21*, 1829–1838.

Hall, R. A., Rappaport, M., Hopkins, H. K., & Griffin, R. Tobacco and evoked potential. *Science*, 1973, *180*, 212–214.

Hauser, H., Schwartz, B., Roth, G., & Bickford, R. Electroencephalographic changes related to smoking. *Electroencephalography and Clinical Neurophysiology*, 1958, *10*, 576.

Heimstra, N. W. The effects of smoking on mood change. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives*. Washington, D.C.: V. H. Winston, 1973.

Heiser, J. F., & DeFrancisco, D. The treatment of pathological panic states with propranolol. *American Journal of Psychiatry*, 1976, *133*, 1389–1394.

Henkin, R. I. The neuroendocrine control of perception. In D. A. Henkin, K. H. Pribram, & A. J. Stunkard (Eds.), *Perception and its disorders: Proceedings of the Association for Research in Nervous Mental Disease*. Baltimore, Md.: Williams & Wilkins, 1970.

Herxheimer, A., Griffiths, R. L., Hamilton, B., & Wakefield, M. Circulatory effects of nicotine aerosol inhalations and cigarette smoking in man. *Lancet*, 1967, *2*, 754–755.

Hill, P., & Wynder, E. L. Smoking and cardiovascular disease—Effect of nicotine on the serum epinephrine and corticoids. *American Heart Journal*, 1974, *87*, 491–496.

Hutchinson, R. R., & Emley, G. B. Effects of nicotine on avoidance, conditioned suppression and aggression response measures in animals and man. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives*. Washington, D.C.: V. H. Winston, 1973.

Ikard, F. F., Green, D. E., & Horn, D. A scale to differentiate between types of smoking as related to the management of affect. *International Journal of the Addictions*, 1969, *4*, 649–659.

Ikard, F. F., & Tompkins, S. The experience of affect as a determinant of smoking behavior: A series of validity studies. *Journal of Abnormal Psychology*, 1973, *81*, 172–181.

Itil, T., Ulett, G., Hsu, W., Klingenberg, H., & Ulett, J. The effects of smoking withdrawal on quantitatively analyzed EEG. *Clinical Electroencephalography*, 1971, *2*, 44–51.

Jarvik, M. E. The role of nicotine in the smoking habit. In W. A. Hunt (Ed.), *Learning mechanisms in smoking*. Chicago: Aldine, 1970.

Jarvik, M. E. Further observations on nicotine as the reinforcing agent in smoking. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives*. Washington, D.C.: V. H. Winston, 1973.

Johnston, L. M. Tobacco smoking and nicotine. *Lancet*, 1942, *2*, 742.

Kershbaum, A., Pappajohn, D. J., & Bellet, S. Effect of smoking and nicotine on adrenocortical secretion. *Journal of the American Medical Association*, 1968, *203*, 275–278.

Knapp, D. E., & Domino, E. F. Action of nicotine on the ascending reticular activating system. *International Journal of Neuropharmacology*, 1962, *1*, 333–351.

Knott, V. J., & Venables, P. H. EEG alpha correlates of nonsmokers, smokers, smoking, and smoking deprivation. *Psychophysiology*, 1977, *14*, 150–156.

Kostowski, W. A note on the effects of some cholinergic and anticholinergic drugs on the aggressive behavior and spontaneous electrical activity of the central nervous system in the ant, *Formica rufa*. *Journal of Pharmacy and Pharmacology*, 1968, *20*, 385–389.

Lacey, J. The evaluation of autonomic response: Toward a general solution. *Annals of the New York Academy of Sciences*, 1956, *67*, 123–164.

Lange, C. G., & James, W. *The emotions*. Baltimore, Md.: Williams & Wilkins, 1922.

Leshner, A. I., et al. Pituitary adrenocortical activity and intermale aggressiveness in isolated mice. *Physiology & Behavior*, 1973, *11*, 705–711.

Levine, S. Stress and behavior. *Scientific American*, 1971, *224*(1), 26–31.

Lindsley, D. B. The role of nonspecific reticulo-thalamo-cortical systems in emotion. In P. Black (Ed.), *Physiological correlates of emotion*. New York: Academic Press, 1970.

Mandler, G. *Mind and emotion*. New York: Wiley, 1975.

Martin, W. R., Sloan, J. W., Sapira, J. D., & Jasinski, D. R. Physiological, subjective, and behavioral effects of amphetamine, methamphetamine, ephedrine, phenmetrazine, and methylphenidate in man. *Clinical Pharmacology and Therapeutics*, 1971, *12*, 245.

Mendenhall, W. L. A study of tobacco smoking. *American Journal of Physiology*, 1925, *72*, 549–557.

Meyers, F. H., Jawetz, E., & Goldfien, A. *Review of medical pharmacology*. Los Altos, Calif.: Lange Medical Publications, 1974.

Miller, N. E. General comments on problems of motivation relevant to smoking. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives*. Washington, D.C.: V. H. Winston, 1973.

Morrison, C. F. The effects of nicotine on punished behaviour. *Psychopharmacologia*, 1969, *14*, 221–232.

Morrison, C. F., & Stephenson, J. A. The occurrence of tolerance to a central depressant effect of nicotine. *British Journal of Pharmacology*, 1972, *46*, 151–156.

Murphree, H. B., Pfeiffer, S., & Price, L. M. Electroencephalographic changes in man following smoking. *Annals of the New York Academy of Sciences*, 1967, *142*, 245–260.

Murphree, H. B., & Schultz, R. Abstinence effects of smokers. *Federation Proceedings*, 1968, *27*, 220.

Myrsten, A. L., Post, B., Frankenhaeuser, M., & Johansson, G. Changes in behavioral and physiological activation induced by cigarette smoking in habitual smokers. *Psychopharmacologia*, 1972, *27*, 305–312.

Nelsen, J. M., & Goldstein, L. Improvement of performance on an attention task with chronic nicotine treatment in rats. *Psychopharmacologia*, 1972, *26*, 347–360.

Nesbitt, P. D. Smoking, physiological arousal, and emotional response. *Journal of Personality and Social Psychology*, 1973, *25*, 137–144.

Nowlis, V. Research with the Mood Adjective Check List. In S. S. Tomkins & C. E. Ikard (Eds.), *Affect, cognition and personality*. New York: Springer, 1965.

Olds, J., & Milner, P. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 1954, *47*, 419–427.

Olds, J., & Olds, M. Drives, rewards and the brain. In F. Barron (Ed.), *New directions in psychology* (Vol. 2). New York: Holt, Rinehart & Winston, 1965.

Philips, C. The EEG changes associated with smoking. *Psychophysiology*, 1971, *8*, 64–74.

Rachman, S. Treatment of prolonged exposure to high intensity stimulation. *Behaviour Research and Therapy*, 1969, *7*, 295–302.

Ray, O. S. *Drugs, society and human behavior*. St. Louis, Mo.: Mosby, 1972.

Roth, G., McDonald, J. B., & Sheard, C. The

effect of smoking cigarettes and the intravenous administration of nicotine on the heart and peripheral blood vessels. *Medical Clinics of North America*, 1945, *29*, 949–957.

Routtenberg, A. The two-arousal hypothesis: Reticular formation and limbic system. *Psychological Review*, 1968, *75*, 51–80.

Satterfield, J. H., Cantwell, D. P., & Satterfield, B. T. Pathophysiology of the hyperactive child syndrome. *Archives of General Psychiatry*, 1974, *31*, 839–844.

Schachter, S. Nesbitt's paradox. In W. L. Dunn (Ed.), *Smoking behavior: Motives and incentives*. Washington, D.C.: V. H. Winston, 1973.

Schachter, S., & Singer, J. E. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 1962, *69*, 379–399.

Schechter, M. D. Effect of nicotine on response to frustrative non-reward in the rat. *European Journal of Pharmacology*, 1974, *29*, 312–315.

Schechter, M. D., & Rand, M. J. Effect of acute deprivation of smoking on aggression and hostility. *Psychopharmacologia*, 1974, *35*, 19–28.

Schildkraut, J. J., & Kety, S. S. Biogenic amines and emotion. *Science*, 1967, *156*, 21–30.

Schuster, C. R. Comments on the paper by Jarvik. In W. A. Hunt (Ed.), *Learning mechanisms in smoking*. Chicago: Aldine, 1970.

Shiffman, S. M., & Jarvik, M. E. Smoking withdrawal symptoms in two weeks of abstinence. *Psychopharmacology*, 1976, *50*, 35–39.

Silverman, A. P. Behavior of rats given a "smoking dose" of nicotine. *Animal Behavior*, 1971, *19*, 67–74.

Silvette, H., Hoff, E. C., Larson, P. S., & Haag, H. G. The actions of nicotine on central nervous functions. *Pharmacological Reviews*, 1962, *14*, 137–173.

Simon, D. L., & Iglauer, A. The acute effect of chewing tobacco and smoking in habitual users. *Annals of the New York Academy of Sciences*, 1967, *142*, 119–132.

Sokolov, E. N. Neuronal models and orienting reflex. In M. A. B. Brazier (Ed.), *Conference on the central nervous system and behavior*. New York: Macy, 1960.

Solomon, R. L., & Corbit, J. D. An opponent-process theory of motivation: II. Cigarette addiction. *Journal of Abnormal Psychology*, 1973, *81*, 158–171.

Ulett, J., & Itil, T. Quantitative electroencephalogram in smoking and smoking deprivation. *Science*, 1969, *164*, 969–970.

Vazquez, A. J., & Toman, J. E. P. Some interactions of nicotine with other drugs on central nervous functions. *Annals of the New York Academy of Sciences*, 1967, *142*, 201–215.

Webster, D. D. The dynamic quantitation of spasticity with automated integrals of passive motion resistance. *Clinical Pharmacology and Therapeutics*, 1964, *5*, 900–908.

Wechsler, R. Effects of cigarette smoking and intravenous nicotine on the brain. *Federation Proceedings*, 1958, *17*, 169.

Wenusch, A., & Schöller, R. Über den Einfluss des Rauchens auf die Reizschwelle des Drucksinnes. *Medizinische Klinik*, 1936, *32*, 356–358.

Wood, D. R., Reimherr, F. W., Wender, P. H., & Johnson, G. E. Diagnosis and treatment of minimal brain dysfunction in adults. *Archives of General Psychiatry*, 1976, *33*, 1453–1460.

Received March 27, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 662–679

History of the Sleeper Effect: Some Logical Pitfalls in Accepting the Null Hypothesis

Thomas D. Cook
Northwestern University

Charles L. Gruder
University of Illinois at Chicago Circle

Karen M. Hennigan and Brian R. Flay
Northwestern University

The history of research on the sleeper effect prior to 1978 can be divided into five stages: (a) initial discovery of the effect, (b) development of the underlying theory, (c) widespread acceptance of the effect and of the discounting cue explanation of it, (d) realization that past operational definitions of the effect were not isomorphic with the conceptual definition, and (e) repeated failures to demonstrate the effect once operational definitions were employed that corresponded to the conceptual definition (Gillig & Greenwald, 1974). These failures resulted in an invitation to accept the null hypothesis and to "lay the sleeper effect to rest." This article illustrates why it is not justifiable to accept the null hypothesis about the sleeper effect. We suggest that provisional acceptance of the null hypothesis depends on assuming that all the necessary theoretical, countervailing, statistical, and procedural conditions for an adequate test of the effect have been demonstrably met. We further suggest that none of the empirical studies prior to 1978 demonstrably succeeded in meeting these conditions. However, adequate tests following the guidelines we have described for provisionally accepting the null hypothesis have recently been conducted, and the effect has been repeatedly found. A deductive model of the logical factors that should guide provisional acceptance of the null hypothesis is contrasted with a current model that stresses induction and statistical power analyses.

Most statistics texts assert that there is no formal basis for accepting the null hypothesis. The rationale for this assumption is basically a restatement of the well-known position in philosophy that inductive knowledge is not logically possible. However, practicing scientists are often forced to act as though the null hypothesis were true even when they know that there is no compelling epistemological basis for their actions. Given the apparent conflict between practice and logic, it is important in the analysis of research findings to be clear about the criteria that help distinguish between a conclusion such as "We can be reasonably certain that no meaningful difference exists" and a conclusion such as "Although no statistically reliable difference was found, we cannot be at all certain whether a meaningful difference exists."

The present article uses the history of research on the sleeper effect to illustrate some pitfalls in accepting the null hypothesis. It also uses this research to describe and justify a model that might help in deciding when no-difference findings warrant the tentative conclusion that no difference exists as opposed to the conclusion that no decision about differences is warranted. The sleeper effect is useful for this illustration, since a recent claim was made that the effect does not exist (Gillig & Greenwald, 1974) despite 25 years of claims that it does. The claim

Karen M. Hennigan is now at the University of Southern California, and Brian R. Flay is now at the University of Waterloo, Waterloo, Canada.
The authors would like to thank Anthony G. Greenwald and Daniel Romer for their comments on a previous draft of this article.
Requests for reprints should be sent to Thomas D. Cook, Department of Psychology, Northwestern University, Evanston, Illinois 60201.

that the effect does not exist was based on a particular model for "How to Accept the Null Hypothesis Gracefully" (Greenwald, 1975, p. 16). However, since subsequent research guided by a quite different model has led to repeated discoveries of the sleeper effect (Gruder, Cook, Hennigan, Flay, Alessis, & Halamaj, 1978), it may be useful to detail the later model here and to contrast it with the earlier model.

To accomplish our ends we first review the history of research on the sleeper effect, for no detailed review yet exists in the literature. Then we outline the requirements for an adequate test of the null hypothesis, using the sleeper effect as an illustration. After this, we examine whether the failures to obtain sleeper effects prior to 1978 were due to the phenomenon's not existing or were due to inadequate tests, and we conclude that no conclusion on this issue was warranted prior to 1978. Next we report the results of two studies that repeatedly obtained sleeper effects when the necessary conditions for the phenomenon were demonstrably met and did not obtain them when these conditions were not met. Finally, we compare and contrast two models for tentatively accepting the null hypothesis.

History of the Sleeper Effect

In 1949, Hovland, Lumsdaine, and Sheffield reported an experiment that was designed to evaluate the impact of a World War II propaganda film on soldiers' beliefs.[1] The experimental design was a 2×2 factorial: Enlisted soldiers did or did not see the film, and their message-relevant beliefs were measured 5 days or 9 weeks afterward. The authors focused their discussion on a subset of eight items in which belief changed little initially and in which the difference between the experimental and control groups was larger after 9 weeks than it was after 5 days. This difference of differences was interpreted as "raising the possibility of a 'sleeper' effect" because it seemed that the full impact of the film took time to build up.

The authors proposed four general hypotheses to explain the effect. The one that has been most cited was called the *discounting cue hypothesis*. It specified (a) that the army, which sponsored the film, was seen as a biased and therefore untrustworthy source for war-relevant information; (b) that its sponsorship of the film led the soldiers to initially discount the filmed message, thereby reducing its immediate impact on their beliefs; (c) that as time passed the source of the message was forgotten or dissociated from the message, thereby removing the change-inhibiting force of the untrustworthy source; and (d) that once the source was no longer linked to the message, soldiers' attitudes rose to the residual level of belief change caused by the message alone. This reasoning suggests a possible explanation of why sleeper effects occurred for only a subset of the belief items in Hovland et al.'s study: The message's initial effectiveness may not have totally dissipated over 9 weeks for these items, but it may have totally dissipated for the other items.

It should be noted that the term *sleeper effect* can be used in a general sense to describe a delayed increase in any dependent variable, not just in belief. Even in persuasion research there is no need to restrict the term to contexts in which the discounting cue hypothesis is invoked as an explanation. Indeed, three of the four explanations that Hovland et al. (1949) offered of their sleeper effect did not even mention a discounting cue. Nonetheless, many commentators in social psychology seem to use the term *sleeper effect* to refer to both the descriptive phenomenon (a delayed increase in belief change) and a particular explanation (the discounting cue hypothesis). We also use the term in this traditional way.

Development of the Theory and Replications of the Sleeper Effect

Three experiments were designed to replicate the sleeper effect and to directly test

---

[1] In the social psychological literature on persuasion, the terms *attitude* and *belief* have often been used interchangeably, although their definitions have been distinguished. Since most studies that we cite measured beliefs, we use this term exclusively in this article, although we recognize that in certain cases attitude may be the appropriate term.

and extend the discounting cue explanation of it. Hovland and Weiss (1951) manipulated whether persuasive messages came from sources of high or low credibility, with low credibility serving as a discounting cue. They claimed that a sleeper effect resulted in the low-credibility condition. Since data indicated that the low-credibility sources had probably not been forgotten, they explained the effect in terms of the message and source being spontaneously dissociated from one another with the passage of time.

Weiss (1953) followed this research by having subjects learn a persuasive message that was or was not linked to a brief statement that discounted the message. Weiss claimed to have found a sleeper effect with this particular discounting cue manipulation, although he carefully pointed out that his inference was based on there being relatively less decay of change when the discounting cue was learned than when it was not. He did not find an absolute increase in belief change in the discounting cue condition.

Kelman and Hovland (1953) had subjects learn a persuasive message from a source of high or low credibility, and the source was or was not reinstated at the delayed testing 2 weeks later. When no reinstatement took place, the data pattern resembled that of Hovland and Weiss: Belief change appeared to increase with time in the low-credibility condition and to decrease with time in the high-credibility condition. But when reinstatement took place (reversing any dissociation that might have occurred) neither the increase in change in low credibility nor the decrease in change in high credibility was obtained.

Kelman and Hovland reasoned from these data that the process of belated dissociation of the message and the cue applies both to cues that cause initial rejection of a message and to cues that enhance a message's initial impact. But while the dissociation of discounting cues should facilitate sleeper effects, the dissociation of message-acceptance cues should accelerate the decay of initial belief change. Kelman and Hovland coined the term *dissociative cue hypothesis* to refer to the common process of first associating and then dissociating message-acceptance or message-rejection cues from a message. Seen from this perspective, the discounting cue hypothesis is a special case of the more general dissociative cue hypothesis.

Widespread Acceptance of the Sleeper Effect

These theory-testing experiments in which sleeper effects were found were followed by many other apparent replications that were reported in the literature from 1953 through the early 1970s. As a result of these repeated replications, numerous textbooks referred both to the validity of the effect and to the validity of the discounting cue explanation of it.

The studies that led to such widespread acceptance of the sleeper effect are reviewed below. We define a discounting cue as any brief item of information that leads a reader to summarily reject the conclusion of a persuasive message and so inhibits the immediate belief change that the message would otherwise cause. The predominant manipulation of a discounting cue has been to have the message come from a communicator of low credibility (Falk, 1970; Gillig & Greenwald, 1974; Hovland & Weiss, 1951; Johnson, Torcivia, & Poprick, 1968; Johnson & Watkins, 1971; Kelman, 1958; Kelman & Hovland, 1953; Schulman & Worrall, 1970; Watts & McGuire, 1964; Weber, 1972; Whittaker & Meade, 1968; Weber, Note 1). But other operationalizations included a brief countercommunication (Weiss, 1953), qualifying statements that questioned the validity of the arguments within the message (Papageorgis, 1963), a countermessage presented once while the persuasive message was repeated three times (Wilson & Miller, 1968), and forewarning that a contrary message existed (Holt & Watts, 1973). We consider all these experiments as relevant to the sleeper effect, though Wilson and Miller and Papageorgis did not specifically mention the discounting cue hypothesis in their reports. (Adding these experiments to the review makes little or no difference to the general conclusions.)

The operational definition of the sleeper effect employed in most studies before the mid-1970s was an interaction of time of test-

ing (immediate vs. delayed) and whether the persuasive message was or was not associated with a discounting cue. For reasons that become clear later, we call this operationalization a relative sleeper effect.

The results of 16 experiments that permitted a test of the relative sleeper effect are presented in columns 2 and 3 of Table 1. Of these studies, 11 resulted in at least marginally significant relative sleeper effects. Thus, the preponderance of the evidence does indicate that the relative sleeper effect is a reliable phenomenon.

However, it should be noted that the absolute values of the slopes of the time trends were typically higher in the experimental conditions without discounting cues than they were in the conditions with them. In other words, decay in the nondiscounting conditions contributed more to the interactions than did any delayed increase in the discounting conditions. This observation raises an important question concerning the fit between the operational and conceptual definitions of the sleeper effect.

Realization That Past Operational Definitions Did Not Match the Conceptual Definition of a Sleeper Effect

Careful reading of early accounts suggests that the crucial defining attribute of a sleeper effect is an absolute increase in belief change over time. For instance, Hovland et al. (1949) summarized their data by stating, "Some of the effects of the film may be 'sleepers' that do not occur immediately but require a lapse of time before the full effect is evident" (p. 188). McGuire (1969) described the same findings as indicating that "the impact of . . . the film was greater after eleven weeks had passed [from the pretest, not the film] than it had been shortly after the showing of the film" (pp. 254–255). In the context of manipulations of low credibility, Insko (1967) wrote, "This increase in the influence of the low credibility source over time was called the 'sleeper effect'" (p. 44).

An operational definition that closely corresponds to the conceptual definition is that there is more belief change at a delayed

Table 1
Outcomes of Past Experiments Classified According to Definition of the Sleeper Effect

	Relative sleeper		Absolute sleeper	
Study	Statistically significant?	Condition with higher slope value	Appropriate direction?	Statistically significant?
Hovland & Weiss (1951)	Yes	Nondiscounting	Yes	DK
Kelman & Hovland (1953)	Yes	Nondiscounting	Yes	No
Kelman (1958)	DK	Discounting	Yes	DK
Watts & McGuire (1964)	Marginal	Nondiscounting	No	
Johnson, Torcivia, & Poprick (1968)	No	Discounting	No	
Whittaker & Meade (1968)	Yes	Nondiscounting	No	
Falk (1970)	Yes	Nondiscounting	Yes	No
Schulman & Worrall (1970)	Yes	Nondiscounting	No	
Johnson & Watkins (1971)	Yes	Nondiscounting	No	
Weber (Note 1)	No	Discounting	No	
Weber (1972)	Yes	Nondiscounting	Yes	Yes
Gillig & Greenwald (1974)	Yes	Nondiscounting	Yes	No
Weiss (1953)	Yes	Nondiscounting	No	
Papageorgis (1963)	No	Nondiscounting	Yes and no	DK
Wilson & Miller (1968)	NA	NA	Yes in 6 of 8 comparisons	DK
Holt & Watts (1973)	Yes	Nondiscounting	Yes	Yes

Note. DK = do not know—cannot be computed from the available data; NA = not appropriate—there was no equivalent to a message-acceptance or message-only condition.

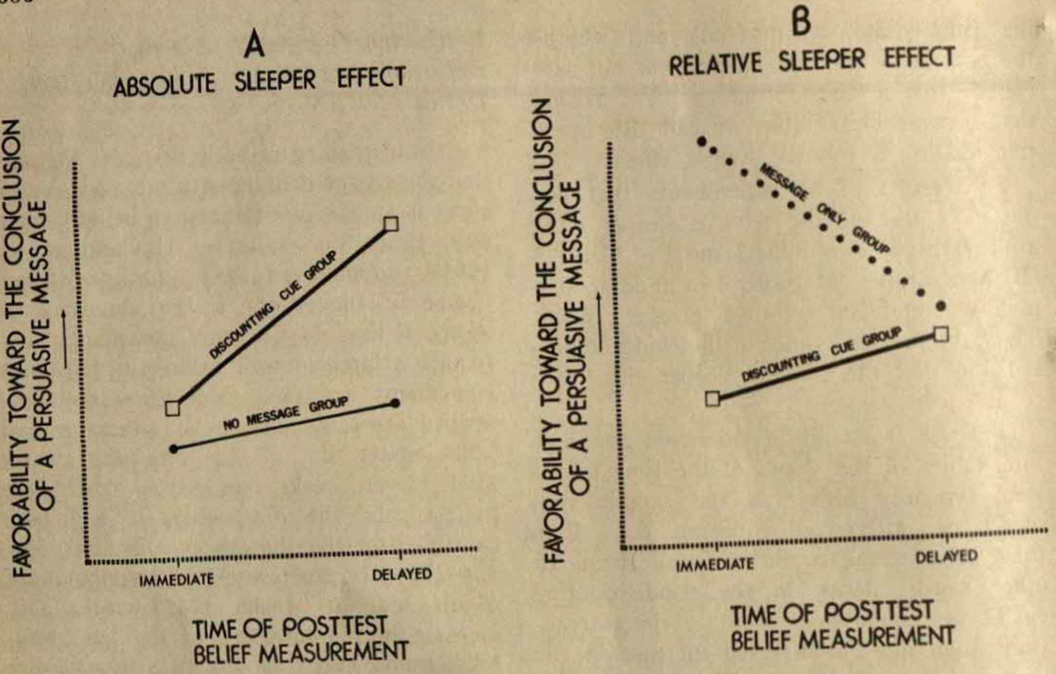

Figure 1. Hypothetical illustrations of absolute and relative sleeper effects.

belief-testing session than at a session immediately after the learning of a message (see the upper line in Figure 1A). However, an implicit assumption of this definition is that beliefs would stay constant over time if there were no persuasive message. To control for the possibility of shifting population beliefs, it is desirable to collect data from no-message controls—subjects who have not received the persuasive message (see the lower line in Figure 1A). A second operational definition of a sleeper effect is that the pattern of posttest beliefs should suggest more of an increase in belief change (or less of a decrease) in a discounting cue group than in a no-message control group. With only two posttest time intervals, a sleeper effect would be indicated by a difference of differences or by an interaction of groups (discounting cue vs. no message) and the time of posttest measurement (immediate vs. delayed). With more delay intervals, differences between time trends would be used for inferring the effect.

In the past, the sleeper effect has often been defined as less decay, or relatively more change over time, in a discounting cue group relative either to a group exposed to the persuasive message alone (message-only controls) or to a group in which the message is linked to a cue that should increase acceptance of the message conclusion (see Figure 1B). These operational definitions rest on the assumption that the decay of initial belief change in the message-only or message-acceptance groups is an accurate estimate of the decay that would result in a discounting cue group if there were no sleeper effect. This assumption is difficult to accept because there is usually more initial change in the comparison groups than in a discounting cue group, and greater subsequent decay is usually associated with large initial changes (Cook & Flay, 1978). Therefore, defining the effect relative to these comparison groups overestimates the expected decay in discounting cue groups, and false positive sleeper effects can be inferred. It is likely that too many false positives were inferred in the past because more initial change was obtained in the groups without discounting cues and because the absolute decay of change found in these groups was more than the absolute increase in change found in the discounting cue groups. In other words, a statistical interaction between time of measure-

ment and experimental groups (a discounting cue group and a message-acceptance or message-only group as the comparison group) was taken as evidence of the effect. These interactions typically resulted from relatively less decay in a discounting cue group rather than from an absolute increase in change, which is what the conceptual definition calls for.

Consideration of the lack of fit between past conceptual and operational definitions of the sleeper effect led Cook (Note 2) to coin the term *relative sleeper effect* to refer to the statistical interaction between experimental groups (one of which receives a message plus a discounting cue, while the other receives the same message without a cue) and time of testing (one test coming immediately after the message and the other after a delay period). Since a relative sleeper effect does not correspond closely to the conceptual definition of a sleeper effect, Cook also coined the term *absolute sleeper effect* to refer to more appropriate operational definitions, namely, a temporal increase in belief change within a discounting cue group or a relatively greater increase in this group when compared with a no-message control group.

There are two special situations in which it is not appropriate to test the absolute sleeper effect in the way we have just outlined. First, the effect would not be validly tested if a discounting cue were so strong that instead of merely inhibiting initial attitude change, it caused attitudes to change significantly in the opposite direction (i.e., it caused a "boomerang effect" to below the no-message baseline). In such cases, one could not distinguish whether any obtained sleeper effect reflected (a) reversion over time toward the no-message baseline, (b) reversion over time toward a level equal to the delayed impact of the message when the message and cue are dissociated, or (c) both forces operating simultaneously. One must, therefore, be careful to inspect immediate belief means lest boomerang effects create spurious sleeper effects or inflate the estimates of true ones. Second, it would be difficult to accept a discounting cue interpretation of an absolute sleeper effect if subjects in a discounting cue group were more favorable to the message conclusion at the delayed posttest than were subjects in a message-only group (i.e., if there were a significant "crossover effect"). This is because the discounting cue hypothesis predicts that after dissociation, belief change in a discounting cue group will increase over time to the level found in a message-only group and not beyond this level. Though it would warrant theoretical and empirical exploration, a significant crossover would be inconsistent with the discounting cue explanation of an obtained absolute sleeper effect.

Failure to Demonstrate the Absolute Sleeper Effect

Prior to 1978, there was only one systematic attempt to test directly for an absolute sleeper effect: Greenwald and Gillig (Note 3) reported their failure to obtain the effect in five experiments and later published an article based on seven unsuccessful replication attempts (Gillig & Greenwald, 1974). In the title of this latter article they rhetorically asked "Is It Time to Lay the Sleeper Effect to Rest?" (Gillig & Greenwald, 1974, p. 132). Their work explicitly acknowledged an earlier version of the present article (Cook, Note 2), in which past experiments were reviewed to see if any conclusions about absolute sleeper effects were warranted. However, no review of this issue has yet appeared in print, and so we briefly offer one here.

There are some difficulties, however, in ascertaining whether past studies resulted in absolute sleeper effects. First, few of the past experiments included statistical tests of the appropriate effect. We have therefore tried to conduct the missing absolute tests ourselves, using estimates of the means and standard errors that were either cited in published reports or that we were able to compute for ourselves from other published information. In three cases, the information was insufficient for computing any direct test of the absolute sleeper effect (Hovland & Weiss, 1951; Kelman, 1958; Wilson & Miller, 1968). Second, some research reports contained relevant data for more than one discounting cue condition (e.g., Hovland and

Weiss had four low-credibility groups), whereas other reports contained data from more than one relevant experiment (e.g., Gillig & Greenwald, 1974). Our strategy in such cases was to examine the data at their lowest level to see if absolute sleeper effects could be tested at that level. Where there were no statistical (or obvious visual) differences across conditions or experiments, we summed over the various discounting cue conditions, and only the summary data are presented here.

Column 4 of Table 1 indicates whether trends in the discounting cue group were in the direction required for an absolute sleeper effect. Column 5 shows whether the effect was statistically significant. Of the 12 experiments with low-credibility manipulations, 6 had trends in the appropriate direction, 4 of which could be statistically tested. None of the trends reached conventional levels of statistical significance. Of the 4 experiments with other discounting cue manipulations, Weiss's (1953) trend was not in the required direction; Papageorgis (1963) obtained the required trend over 14 days but not over 41 (the 14-day trend could not be statistically evaluated); Wilson and Miller (1968) seem to have obtained the appropriate trends in some of their conditions, but statistical evaluation was again impossible; Holt and Watts (1973) obtained a statistically significant effect.

All in all, there is still no convincing evidence of absolute sleeper effects where low source credibility was the discounting cue. The evidence looks more promising with the other manipulations, but two facts have to be remembered about them. First, the Wilson and Miller experiment was not conceived as relevant to the discounting cue hypothesis; we have made the connection. Second, Holt and Watts found a marginal trend that indicated that subjects in the discounting cue group recalled more of the message both immediately after its presentation and 1 week later. Although the absence of a simple relationship between message learning and the persistence of attitude change (Cook & Flay, 1978) suggests that learning factors might not have mediated the absolute sleeper effect that these authors obtained, it would be more comforting if the discounting cue had not been confounded with marginally greater initial learning and subsequent retention of the message.

In the face of the preponderance of evidence from the past experiments reviewed here and their own seven failures to find an absolute sleeper effect, Gillig and Greenwald (1974) invited readers to accept the null hypothesis that there is no (absolute) sleeper effect. This invitation has not gone unheeded, and recent textbooks have begun to reflect skepticism about the validity of the sleeper effect (e.g., Baron & Byrne, 1977; Oskamp, 1977; Schneider, 1976). To assess whether it is warranted to accept the null hypothesis in this instance, we first discuss the logical requirements for accepting the null hypothesis and then apply this logic to the case of the sleeper effect.

Logic of Provisionally Accepting the Null Hypothesis

Statistics texts teach that it is logically impossible to accept the null hypothesis. However, practical concerns demand that one sometimes provisionally act as though the null hypothesis were true. We believe that there are some guidelines that help in determining when no-difference findings warrant tentative acceptance of the null hypothesis. These are as follows: (a) when the theoretical conditions necessary for the effect to occur have been explicated, operationalized, and demonstrably met in the research; (b) when all the known plausible countervailing forces have been explicated, operationalized, and demonstrably ruled out; (c) when the statistical analysis is powerful enough to detect at least the theoretically expected maximum effect at a preordained alpha level; and (d) when the manipulations and measures are demonstrably valid.

Theoretical Conditions Necessary for the Effect to Occur Must Be Met

Scientific propositions about the existence of a phenomenon are not adequately tested if the theoretical conditions necessary for the phenomenon are absent from empirical

tests. In the case of the sleeper effect, one therefore has to explicate the underlying theory—the discounting cue hypothesis—to determine when it predicts that absolute sleeper effects should occur.

Examination of the discounting cue hypothesis reveals that sleeper effects are only expected under certain conditions. These conditions can perhaps best be understood by considering Figure 2. It shows belief at three different posttest time intervals (immediate and two different delays) in a discounting cue group, a message-only group, and a no-message control group. The figure has been drawn to illustrate the case in which (a) the message causes initial belief change, (b) this change either decays at a continuous rate (the dashed line) or does not decay at all (the dotted line), (c) the discounting cue suppresses all initial belief change (point *A*); and (d) population beliefs do not shift over time (see the solid line for the no-message control group).

Consider what would happen if the message and the discounting cue became dissociated some time before Posttest 2 (delayed). Imagine first that there was no decay of change in the message-only condition (the dotted line). Then, the theory predicts that after dissociation, belief in the discounting

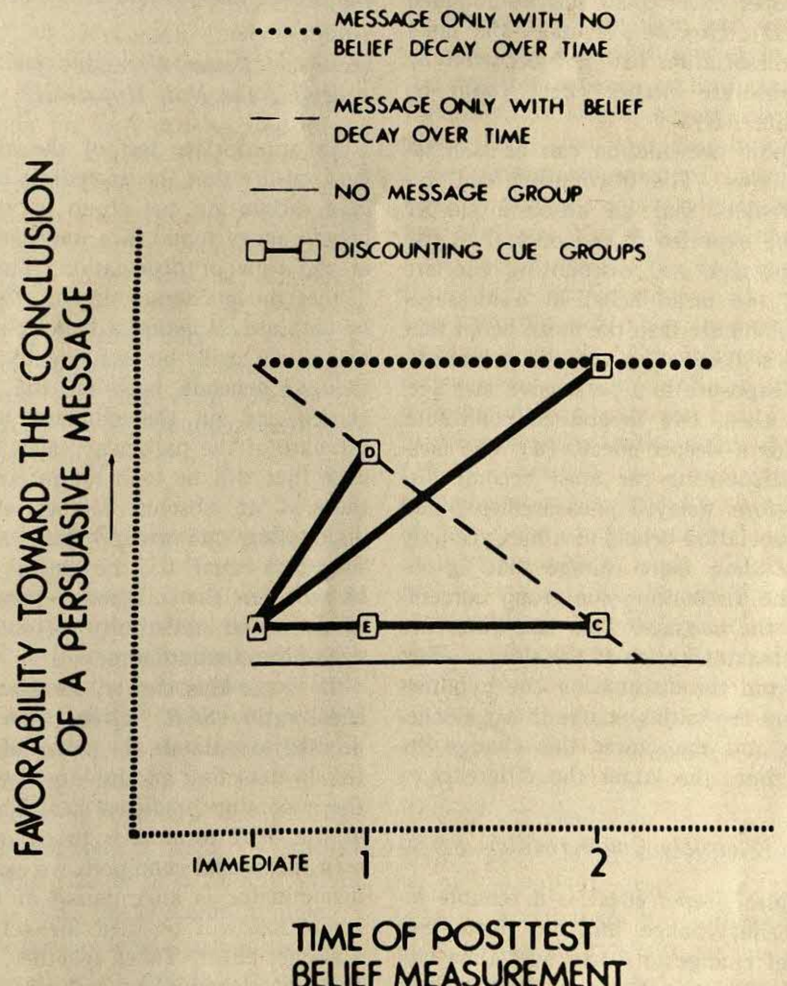

Figure 2. Hypothetical time trends that illustrate that an absolute sleeper effect is a special case of the discounting cue hypothesis.

cue condition will increase to the level found in the message-only condition at the time of measurement. As line *AB* shows, an absolute sleeper effect would be expected to result under these conditions. Imagine next what would happen if all the conditions were the same except that there was decay of belief change in the message-only condition (the dashed line). Then, no absolute sleeper effect would be expected to occur (see line *AC*), though everything else was the same as for line *AB*.

Consider next what would happen if beliefs in the message-only condition were decaying (the dashed line) and dissociation had taken place by Posttest 1 (delayed). An absolute sleeper effect could still be obtained (see line *AD*). However, if under the same conditions dissociation had not occurred by then, no absolute sleeper effect could be found (see line *AE*).

This graphic presentation can be summarized as follows: The discounting cue hypothesis predicts that an absolute sleeper effect can be expected if and only if at the time the message and discounting cue are dissociated, the mean belief in a message-only group is higher than the mean belief that is found in a discounting cue group immediately after exposure to a persuasive message. There are, then, two theoretical conditions necessary for a sleeper effect: (a) The message and discounting cue must become dissociated before delayed measurement, and (b) postdissociation beliefs in a message-only group must show more change than is obtained in the discounting cue group immediately after the message. This last difference defines the maximum size of the sleeper effect predicted from the discounting cue hypothesis; the more the initial change in a message-only group and the slower this change decays with time, the larger the difference.

Ruling Out Plausible Countervailing Forces

An absolute sleeper effect is a reliable increase in belief change, and any force that causes belief change to decay over time will countervail against such an effect. It is commonplace in persuasion research to note that experimentally induced belief changes decay with time. In their review of 30 years of relevant research, Cook and Flay (1978) concluded that despite particular instances of total persistence or sleeper effects attributable to forces other than the discounting cue hypothesis, decay of initial change was the modal finding. Therefore, it is important in designing adequate tests of the sleeper effect to ensure that decay forces of all kinds are minimized. The countervailing power of belief decay is minimal in experiments in which the discounting cue suppresses all of the belief change a message would otherwise have caused. However, such suppression cannot be due to chance, for if it were, statistical regression would result and would masquerade as a sleeper effect.

Statistical Power Necessary for Accepting the Null Hypothesis

An appropriate test of the sleeper effect must ensure that the immediate belief change in a dicounting cue group is less than the belief change found in a message-only group at the time of dissociation. This difference defines the maximum size of effect that can be obtained. Whether a belief change of this magnitude will be statistically significant, though, depends both on the alpha level chosen and on the standard error of the estimate of the particular mean belief difference that will be used for inferring whether there is an absolute sleeper effect in the discounting cue group. Once alpha is assumed to equal .05, the crucial determinant of whether the maximum possible sleeper effect can be statistically corroborated is the size of the standard error.

To make this clearer, consider the sleeper effect ratio (*SER*), given below, which can be used to estimate the power of a statistical test in detecting an absolute sleeper effect of the maximum predicted size. The numerator of the *SER* formula is based on the necessary theoretical conditions we explicated. The denominator is an estimate of the standard error that will be used for actually testing a sleeper effect. Taken together, the numerator and denominator indicate whether the theoretically specified maximum distance over which belief can increase over time in

a discounting cue group is large enough so that if a sleeper effect were to be obtained it could be statistically corroborated. The minimal requirement for an adequate test of the sleeper effect is that *SER* exceed the *t* value associated with the alpha level and the degrees of freedom that are to be used in direct tests of the absolute sleeper effect. The formula is as follows:

$$SER = \frac{M_{\mathrm{PD \cdot MO}} - M_{\mathrm{IMM \cdot DC}}}{\text{Particular standard error used for directly testing the absolute sleeper effect}},$$

where $M_{\mathrm{PD \cdot MO}}$ is the belief mean at the post-dissociation testing (PD) in a message-only group (MO), and $M_{\mathrm{IMM \cdot DC}}$ is the belief mean at the immediate postmessage testing (IMM) in a discounting cue group (DC).

The formula for *SER* can be used in any study to inquire whether the theoretically specified necessary conditions for an adequate test of the sleeper effect are met and whether a statistical test powerful enough to detect the effect, if it exists, can be conducted. However, in practice *SER* can somewhat overestimate or underestimate the power of the subsequent test of the absolute sleeper effect, because sampling error makes the sample means in the *SER* numerator deviate from their population values.

Valid Manipulation and Measurement

Of all the necessary conditions for an adequate test of a phenomenon, perhaps the most obvious is that it must be shown that the manipulation of the independent variable was effective and that the dependent variable was validly measured. In the context of the discounting cue hypothesis, this amounts to finding differences in immediate belief change across groups that were or were not exposed to a discounting cue. Such a finding would illustrate (a) that the discounting cue is effective in reducing belief change and (b) that a measure of belief with face validity is at least reliable enough to discriminate between group means.

Did Past Tests Meet the Logical Requirements for an Adequate Test of the Sleeper Effect?

Was the Countervailing Decay Force Ruled Out?

An adequate test of the sleeper effect requires that there be no immediate belief change in the discounting cue group. The third column of Table 2 shows that immediate belief change occurred in all but three of the experiments in which the relevant differences could be tested. Thus, in most past experiments a decay force may have been set up in the discounting cue group that may have countervailed against obtaining an absolute sleeper effect and may have led to less than adequate tests of it.

Does the evidence confirm our contention that absolute sleeper effects are more likely when there is no immediate belief change in a discounting cue group? To answer this, we need to contrast the results in which there was and was not immediate change. Hovland and Weiss (1951) had two messages that caused immediate attitude change and two that did not. Over 4 weeks, the percentage of subjects with promessage attitudes increased by 30% and 27% for the messages without initial change and by 13% and 3% for the messages with initial change. Falk (1970) obtained no immediate attitude change, and in comparison with other studies, he found one of the strongest indications of an absolute sleeper effect, albeit nonsignificant ($t = 1.50$). Finally, Wilson and Miller (1968) found no initial belief change in the eight conditions in which prosecution arguments preceded defense ones, but found initial change in the eight other conditions in which the order of arguments was reversed. There were some trends in the sleeper effect direction in conditions in which there was no immediate belief change, and there were no indications of sleeper effects in conditions in which there was immediate change.

Ex post facto evidence of this kind is by no means definitive, but it is consistent with our suggestion that absolute sleeper effects are more likely when no immediate belief

Table 2
Extent to Which Requirements for a Strong Test of the Absolute Sleeper Effect Were Met in Past Experiments

				Proxies of the extent to which dissociation of the message and cue took place		
Study	Condition	Was there initial change in discounting cue condition?	Were the requirements of the *SER* formula met?	Percentage recalling cue	Final delay period (in weeks)	Did discounting cue mean differ from others at delayed test?
Hovland & Weiss (1951)	Some messages	Yes	No	55–77	4	No
	Other messages	No	No		4	No
Kelman & Hovland (1953)		DK	No	DK	3	No
Kelman (1958)		DK	DK	DK	1.5	DK
Watts & McGuire (1964)	1–2 weeks	Yes	Yes	36	1–2	Yes
	6 weeks	Yes	No	10	6	No
Johnson, Torcivia, & Poprick (1968)		Yes	Yes	DK	1	Yes
Whittaker & Meade (1968)		Yes	No	DK	4	No
Falk (1970)		No	No	DK	3.9	No
Schulman & Worrall (1970)		Yes	No	8	3.7	No
Johnson & Watkins (1971)	Message 1 time	Yes	No	DK	4	No
	Message 5 times	Yes	Yes	DK	4	Yes
Weber (Note 1)		Yes	No	DK	4	No
Weber (1972)	3 weeks	Yes	Yes	2[a]	3	Yes
	7 weeks	Yes	No		7	No
Gillig & Greenwald (1974)		Yes	Yes	DK	2	Yes
Weiss (1953)		Yes	No	6	6	No
Papageorgis (1963)	2 weeks	Yes	Yes	"Low" recall[b]	2	Yes
	6 weeks	Yes	No		6	Yes
Wilson & Miller (1968)	Prosecution–defense order	No	DK	DK	1	DK
	Defense–prosecution order	Yes				
Holt & Watts (1973)		DK	Yes	DK	1	No

Note. SER = sleeper effect ratio. DK = do not know—cannot be computed from available data.
[a] This figure may not be a true index inasmuch as the data were gathered in an entirely different context from the session in which the source was introduced. The experimenter, setting, and ostensible purpose of the delayed session were intentionally divorced from the first session to minimize the effects of memory and associative cues.
[b] Based on one of the two experimental messages only where an average of .25 qualifiers were recognized out of a total of 2. Chance responding would give .50 correct.

change occurs in a discounting cue condition than when it does.

Were the Necessary Theoretical Conditions Met in Tests With Sufficient Statistical Power?

In its numerator, the *SER* formula incorporates the necessary theoretical condition that the impact of the message at the time of dissociation be greater than the immediate impact of the message when it is linked to a discounting cue. Via its denominator, the *SER* formula probes whether in a particular experiment, sufficient statistical power exists to obtain a sleeper effect of the maximal predicted value. Theoretically, *SER* can be applied to past studies to determine the extent to which each constituted a powerful test of the absolute sleeper effect. In practice, *SER* is not as useful as it could be, since it presupposes the measurement of dissociation, and measures of dissociation are rare in past studies. Moreover, where available, they are invariably associated with single measures that have particular biases. For instance, Watts and McGuire (1964) measured subjects' recognition of the source's name, which presumably overestimates both recall and association. As another example, Schulman and Worrall (1970) had subjects free associate to the message topic, a procedure that may underestimate association, since some subjects may think it inappropriate to mention a source when asked to free associate to a topic. Although it is much better to have such measures than to be without them, they are not by themselves particularly good indicators of message–cue dissociation.

The lack of data on dissociation in past studies forced us to estimate a modified form of *SER*. Since we did not know when dissociation reached acceptable levels, we computed a separate *SER* value for each time interval at which belief measurement took place. Hence, for each study, there were as many *SER* values as delay intervals. To aid interpretation we assumed that dissociation is positively related to the length of the delay interval and is approximately indexed by the percentage of subjects who do not report recognizing or spontaneously recalling a discounting cue.

For only seven studies did the modified *SER* values we computed exceed the t value associated with the statistical test of the absolute sleeper effect, and this occurred in only some experimental groups: (a) in the low-credibility group of Watts and McGuire (1964) after 1 and 2 weeks but not after 6; (b) in Johnson et al. (1968); (c) in the low-credibility–message-five-times condition of Johnson and Watkins (1971), but not in the condition in which the message was presented once; (d) in the experiments of Gillig and Greenwald (1974); (e) in the 3-week-delay condition of Weber (1972) in which the source was mentioned twice, but not in the 7-week-delay condition; (f) in the 2- and 14-day-delay conditions, but not the 41-day-delay condition, of Papageorgis (1963); and (g) for the minor conclusion measure of Holt and Watts (1973).

Two points are worth noting about these experiments. First, immediate belief change occurred in all the discounting cue groups, with the only possible exception being Holt and Watts, in whose study the relevant test of immediate change could not be conducted. Consequently, there are no experiments in which there was both an absence of initial change and a modified *SER* value greater than the t value associated with the statistical test of an absolute sleeper effect. Second, the data suggest that little dissociation may have occurred in studies with higher modified *SER* values. Consider the last three columns of Table 2. Of Watts and McGuire's (1964) subjects, 36% still recognized the low-credibility source after 1 and 2 weeks. Johnson et al. (1968) and Holt and Watts (1973) had a final delay period of only 1 week, and Gillig and Greenwald (1974) had a final delay of only 2 weeks. The typical decay of belief change did not occur in the high-credibility–message-five-times conditions of Johnson and Watkins (1971), which suggests that the message and source may not have been dissociated (although other explanations are possible). And the modified *SER* values in Weber (1972) and Papageorgis (1963) were only large enough with the shorter delay intervals (i.e., at 3 weeks but

not 7 for Weber and at 2 weeks but not 6 for Papageorgis).

The evidence we have just reviewed about dissociation and *SER* values is at best suggestive, but it does imply that high levels of dissociation may not have been obtained in most past experiments in which the modified *SER* value was large enough for a powerful statistical test. More important, it should be noted that all of the experiments with high *SER* values were characterized by immediate belief change in the discounting cue condition. Since a strong test of the sleeper effect requires both the absence of initial change in the discounting cue group and a large enough *SER* value, it is not clear whether the absence of past convincing demonstrations of the sleeper effect was due to the unreliability of the phenomenon or to weak tests of it.

Do Adequate Tests Detect the Sleeper Effect?

Reasons to Suspect That They Do

The results prior to 1978 of tests of the absolute sleeper predicted from the discounting cue hypothesis are ambiguous because one could conclude from them either that there is no absolute sleeper effect or that the tests of it were inadequate. There are two major reasons for suspecting that the latter is true and that the effect might be found if appropriately tested. First, our review of past studies shows that past failures to obtain a sleeper effect occurred in studies in which it is unlikely (a) that all of the theoretically relevant conditions for a strong test were met and (b) that the countervailing force was ruled out. Moreover, ex post facto analyses suggested that the less the initial belief change in the discounting cue group, the stronger the trends toward a sleeper effect.

Second, the dissociative cue hypothesis, the basic theory from which the sleeper effect is derived, was twice directly tested, and each time was corroborated. The dissociative cue hypothesis proposes that both message-acceptance and message-rejection cues can be dissociated from a message over time and that this dissociation eliminates the impact that the cues initially have on belief. Thus, the dissociative cue hypothesis predicts that when a message-acceptance cue is dissociated the decay of change will be accelerated but that when a message-discounting cue is dissociated the initial inhibition of belief change will be eliminated and an absolute sleeper effect will sometimes result. In this sense, the dissociative cue hypothesis is a more general statement of the discounting cue hypothesis we have examined in this article.

Direct tests of the dissociative cue hypothesis have varied the degree of message–source association. Kelman and Hovland (1953) did this when the original message source (of high or low credibility) was or was not reinstated at the delayed measurement session. As predicted from the dissociative cue hypothesis, (a) in high-credibility conditions, attitude decayed less when the source was reinstated than when it was not, and (b) in low-credibility conditions, belief showed a trend toward a sleeper effect when the source was not reinstated and no such trend when it was. In a related experiment, Weber (1972) manipulated the strength of the message–cue association during exposure to the persuasive message rather than by reinstating the source at the delayed measurement. Weber had subjects learn about a high- or low-credibility communicator whose name was mentioned twice (low association) or 22 times (high association) in the course of reading and then rereading a long persuasive message. Belief was assessed immediately after the message for all subjects, and 3 weeks later for some of them and 7 weeks later for the remainder. The delayed assessments were conducted in subjects' living quarters by experimenters who did not know the hypothesis, and it is likely that the earlier experiment was not reinstated. Weber's time trends (see Figure 3) are very similar to those of Kelman and Hovland, and the overall three-way interaction of credibility, strength of association, and time of measurement was statistically significant at the 5% level (one-tailed test). It seems, then, that the basic dissociative cue hypothesis is correct. This hypothesis buttresses the prediction of a sleeper effect,

which suggests—but does not prove—that such an effect would be obtained if a strong test were conducted.

Results of Adequate Tests of the Sleeper Effect

When demonstraby adequate tests of the absolute sleeper effect were conducted, they repeatedly resulted in absolute sleeper effects (Gruder et al., 1978). In one experiment subjects read a persuasive message about either the disadvantages of a 4-day work week or the disadvantages of permitting right turns when traffic lights are red. Each message was or was not accompanied by a brief statement that suggested that the information in the message was false. Subjects' beliefs were assessed both immediately after the message and 5 weeks later at an unrelated experimental session. The data showed (a) that the discounting cue manipulation suppressed all belief change for the 4-day work week topic but not for the right turn on red topic and (b) that *SER* was large enough for the 4-day work week topic but not for the right turn on red topic. Thus, the requirements for an adequate test of the absolute sleeper effect were clearly met for one message and not for the other. Analysis of the belief data showed a statistically significant sleeper effect for the message in which the requirements were met but no such effect for the message in which they were not met.

A second experiment was designed to replicate and extend the first by linking the 4-day work week message to five different discounting cues that varied in strength and kind. Half the subjects' beliefs were measured immediately after the message, and all the subjects' beliefs were measured by phone 6 weeks later by persons who did not know the subjects' experimental conditions. Analyses of the data relevant to the requirements for an adequate test showed that (a) there was no immediate attitude change in three of the five discounting cue groups; (b) *SER* was large enough for these three groups; and (c) direct measures of dissociation revealed that over 75% of the subjects in each cue group dissociated or forgot the cue. Thus, the requirements for an adequate test were clearly met for three cue groups, but were not met for the other two. Absolute sleeper effects were reliably obtained in the three cue groups in which the requirements were met and not in the two cue groups in which the requirements were not met.

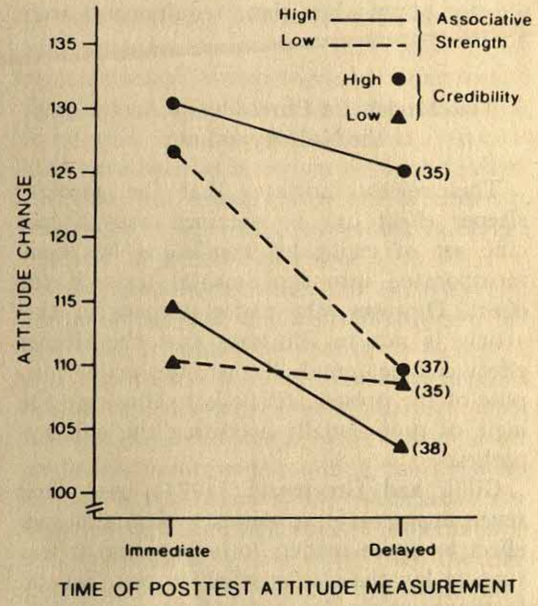

Figure 3. Persistence of attitude change as a function of source credibility and source–content associative strength in Weber (1972), whose relevant data are reproduced with his permission. (Sample sizes are in parentheses, and the 3- and 7-week-delay groups are collapsed because there were no theoretically meaningful effects of the length of the delay period.)

These studies demonstrate that the absolute sleeper effect that is predicted from the discounting cue hypothesis is both reliable and valid. It is reliable in the sense that it has been replicated across experiments, discounting cues within experiments, repeated versus not repeated belief measurements, and two different techniques of delayed measurement. Though replication with different messages and in different laboratories is still needed, it does seem that the absolute sleeper effect is valid by criteria of both convergent and divergent validity insofar as the effect occurred when the logical requirements for an adequate test were demonstrably met and

did not occur when these requirements were not met.

Two Models for Provisionally Accepting the Null Hypothesis

This review indicates that the absolute sleeper effect can be obtained once a specific set of explicated conditions has been incorporated into experimental tests of the effect. However, the major purpose of this article is not to illustrate that the sleeper effect can be found. Rather, the major purpose of the present article is to illustrate the logic of provisionally accepting the null hypothesis.

Gillig and Greenwald (1974) used their seven unsuccessful attempts to find a sleeper effect to invite readers to believe that it was time to lay the sleeper effect to rest. Several textbook writers have already accepted their invitation, despite the fact that it may well be based on an implicit inductive fallacy: The more numerous the failures to obtain an effect, the less likely it is to exist. In a later article, Greenwald (1975) used the so-called nonexistence of sleeper effects as a major illustration of the factors that indicate "How to Accept the Null Hypothesis Gracefully" (p. 16). These factors are as follows:

1. "Use a range, rather than a point, null hypothesis" (p. 16).
2. "Select N on the basis of a desirable error of estimate of the test statistic" (p. 16).
3. "Have convincing evidence that manipulations and measures are valid" (p. 17).
4. "Compute the posterior probability of the null (range) hypothesis" (p. 17). (Doing this led Greenwald to conclude from his own studies of the sleeper effect that the posterior odds ratio in favor of the null hypothesis is 1:249 as opposed to 1:19 for $\alpha = .05$.)
5. "Report all results of research for which conditions appropriate to testing a given hypothesis have been established" (p. 18). (Greenwald elaborated on this by stressing the Type I errors that result from the nonpublication of nonsignificant findings.)

With the exception of Point 3, Greenwald's list is statistical, and the casual reader might infer from this and from Gillig and Greenwald that acceptance of the null hypothesis is largely a matter of the statistical power of a study or a set of related studies and that the more past failures to obtain the effect in question, the higher should be the confidence in accepting the null hypothesis. Although this perspective on the null hypothesis is useful, it must be incomplete, since it led to incorrect acceptance of the null hypothesis about the sleeper effect. What is incomplete about this perspective?

First, it does not stress the importance of deducing from the parent theory (the discounting cue hypothesis in the sleeper effect case) the theoretical conditions that are necessary for a particular effect. Explication of the discounting cue hypothesis makes it clear that the effect should only be predicted when the belief found in a discounting cue group immediately after a message is at a lower level than the belief found in a message-only group just after dissociation. Given that the parent theory specifies the conditions under which a sleeper effect should be obtained, it is incumbent on any researcher to ensure that these conditions are met in experimental tests. If the conditions are not met, it does not matter how many tests are conducted, because they will all be inadequate.

Second, it is important to use background theory and common sense to specify extraneous forces that might plausibly countervail against the phenomenon under investigation. In the sleeper effect case, one very plausible countervailing force is the decay of belief change. Such decay is regularly found after belief change has been obtained as a consequence of an effective persuasive message. Of course, belief does not invariably decay, and it is logically impossible to specify all the potential suppressor variables. But despite such complications, researchers who want to provisionally accept the null hypothesis need to show that countervailing forces have not operated in studies to mask the effect under investigation.

Third, statistical considerations are crucial if one wants to act as though the null hypothesis were true. This is clearly recognized

in Greenwald's (1975) perspective, which urges researchers to design experiments so that the standard error of estimate of the test statistic is desirable. However, how should one define what is desirable? Fortunately, there are some cases in which theory allows one to specify a numerical value for the desirable standard error. In the sleeper effect case, the maximum possible size of the effect can be specified by subtracting the difference between the immediate posttest belief mean in a discounting cue group from the postdissociation belief mean in a message-only group. Then one can easily solve for the size of the standard error that is necessary for an adequate test of the sleeper effect. (The *SER* formula incorporates all of the factors necessary for the computation in question.) There are, of course, many situations in which the underlying theory is not so precise and in which the practicing researcher does not have independent knowledge of the maximum magnitude of the effect. In such situations, one might well follow Greenwald's advice and keep the standard error low by using range estimates of the effect instead of points. But where the theory specifies the magnitude of an expected effect, the size of the necessary standard error can be computed.

On another issue our perspective entirely overlaps with Greenwald's, for accepting the null hypothesis depends on demonstrating that the manipulations and measures are valid.

The essence of the difference between the two perspectives is that Greenwald's stresses statistical concerns pertinent to evaluating no-difference results from a single study or a set of highly similar studies, whereas our perspective stresses the logic of using theory prior to data collection to deduce the conditions that are necessary for an effect and to deduce any conditions that might countervail against the effect. Unless these conditions are explicated and demonstrably incorporated into the research, statistical criteria matter little, and the number of failures to obtain a given effect is uninformative. Our stress, then, is (a) on the logic of explicating the necessary theoretical, countervailing, statistical, and procedural conditions for an effect; (b) on deducing which research operations will represent these conditions; and (c) on demonstrating via appropriate measurement that all of these conditions have been incorporated into the experimental tests. If they have been so incorporated but the effect still does not appear, then one is closer to being able to act with justification as though the null hypothesis were true.

Other factors, although not necessary for adequate tests of the null hypothesis, are desirable. It would be foolish to accept a null hypothesis if some secondary implication of this acceptance were patently false. For instance, the sleeper effect is a single prediction from the discounting cue hypothesis that is itself part of the dissociative cue hypothesis. Since independent experimental tests have strongly suggested that the dissociative and discounting cue hypotheses may well be correct (Kelman & Hovland, 1953; Weber, 1972), this alone should give one reason to pause before using no-difference findings to claim that the sleeper effect should be laid to rest.

When accepting the null hypothesis, there is sometimes value in replication. Cronbach and Snow (1977) have correctly pointed out that if many unbiased studies show nonsignificant effects in the same direction, then the null hypothesis can be rejected. (The stress here has to be on *unbiased,* for multiple biased tests provide no information if the bias operates in the same direction across all the studies.) By the same token, acceptance of the null hypothesis is facilitated if in many studies in which the necessary logical conditions for an effect have been met, the effect does not appear and there is no pattern to the nonsignificant results.

Other advantages of replication emerge in considering the question, Are there any circumstances under which one could incorporate all the necessary theoretical, countervailing, statistical, and procedural requirements into an experiment but still fail to find a true effect? Obviously, chance dictates that a Type II error can occur, and replication is the way to examine this threat. A less obvious circumstance in which one might falsely conclude in favor of the null hypothesis is one in which all the known necessary condi-

tions for an effect are met but unknown necessary conditions are not. Imagine that absolute sleeper effects can occur under the conditions we have specified, but only if the discounting cue follows the persuasive message. In this case, experimental tests that met all the derived necessary conditions but presented the discounting cue before the message would lead to a false acceptance of the null hypothesis. The practical difficulty with such unknown conditions is that they cannot be deliberately incorporated into the research. Consequently, it is advantageous to postpone acting as though the null hypothesis were true until a number of studies have been conducted that meet all the necessary conditions derived from the theory and also vary a number of factors that might plausibly be additional necessary conditions.

Since one can never know in practice whether one has met the true necessary (and sufficient) conditions for any effect, inferences about the null hypothesis must inevitably and logically be imperfect and subject to later correction. As Cook and Campbell (1979) have pointed out in their extended discussion of the null hypothesis, inferences in favor of the null hypothesis should be thought of as decisions to act as though this hypothesis were true and not as knowledge that the hypothesis is true.

Reference Notes

1. Weber, S. J. *Source primacy-recency effects and the sleeper effect.* Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1971.
2. Cook, T. D. *The discounting cue hypothesis and the sleeper effect.* Unpublished manuscript, Northwestern University, 1971. (Available from Thomas D. Cook, Department of Psychology, Northwestern University, Evanston, Illinois 60201.)
3. Greenwald, A. G., & Gillig, P. M. *A cognitive response analysis of the "sleeper effect."* Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1971.

References

Baron, R. A., & Byrne, D. *Social psychology: Understanding human interaction* (2nd ed.). Boston: Allyn & Bacon, 1977.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and analysis issues for social research in field settings.* Chicago: Rand McNally, 1979.

Cook, T. D., & Flay, B. R. The temporal persistence of experimentally induced attitude change: An evaluative review. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11). New York: Academic Press, 1978.

Cronbach, L. J., & Snow, R. E. *Aptitudes and instructional methods.* New York: Irvington, 1977.

Falk, D. I. *The effects on attitude change of manipulating antecedents of Kelman's internalization process.* Unpublished master's thesis, Northwestern University, 1970.

Gillig, P. M., & Greenwald, A. G. Is it time to lay the sleeper effect to rest? *Journal of Personality and Social Psychology,* 1974, *29,* 132–139.

Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychological Bulletin,* 1975, *82,* 1–20.

Gruder, C. L., Cook, T. D., Hennigan, K. M., Flay, B. R., Alessis, C., & Halamaj, J. Empirical tests of the absolute sleeper effect predicted from the discounting cue hypothesis. *Journal of Personality and Social Psychology,* 1978, *35,* 1061–1074.

Holt, L. E., & Watts, W. H. Immediate and delayed effects of forewarning of persuasive intent. *Proceedings of the 82nd Annual Convention of the American Psychological Association,* 1973, *8,* 361–362. (Summary)

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. *Experiments on mass communication.* Princeton, N.J.: Princeton University Press, 1949.

Hovland, C. I., & Weiss, W. The influence of source credibility on communication effectiveness. *Public Opinion Quarterly,* 1951, *15,* 635–650.

Insko, C. A. *Theories of attitude change.* New York: Appleton-Century-Crofts, 1967.

Johnson, H. H., Torcivia, J. M., & Poprick, M. A. Effects of source credibility on the relationship between authoritarianism and attitude change. *Journal of Personality and Social Psychology,* 1968, *9,* 179–183.

Johnson, H. H., & Watkins, T. A. The effects of message repetition on immediate and delayed attitude change. *Psychonomic Science,* 1971, *22,* 101–103.

Kelman, H. C. Compliance, identification, and internalization: Three processes of opinion change. *Journal of Conflict Resolution,* 1958, *2,* 51–60.

Kelman, H. C., & Hovland, C. I. Reinstatement of the communicator in delayed measurement of opinion change. *Journal of Abnormal and Social Psychology,* 1953, *48,* 327–335.

McGuire, W. J. The nature of attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 3, 2nd ed.). Reading, Mass.: Addison-Wesley, 1969.

Oskamp, S. *Attitudes and opinions.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Papageorgis, D. Bartlett effect and the persistence

of induced opinion change. *Journal of Abnormal and Social Psychology,* 1963, *67,* 61–67.

Schneider, D. J. *Social psychology*. Reading, Mass.: Addison-Wesley, 1976.

Schulman, G. I., & Worrall, C. Salience patterns, source credibility, and the sleeper effect. *Public Opinion Quarterly,* 1970, *34,* 371–382.

Watts, W. A., & McGuire, W. J. Persistence of induced opinion change and retention of the inducing message content. *Journal of Abnormal and Social Psychology,* 1964, *68,* 233–241.

Weber, S. J. Opinion change as a function of the associative learning of content and source factors. (Doctoral dissertation, Northwestern University, 1972). *Dissertation Abstracts International,* 1972, *33,* 2798A. (University Microfilms No. 72-32,607)

Weiss, W. A "sleeper" effect in opinion change. *Journal of Abnormal and Social Psychology,* 1953, *48,* 173–180.

Whittaker, T. O., & Meade, R. D. Retention of opinion change as a function of differential source credibility. *International Journal of Psychology,* 1968, *3,* 103–108.

Wilson, W., & Miller, H. Repetition, order of presentation, and timing of arguments and measures as determinants of opinion change. *Journal of Personality and Social Psychology,* 1968, *9,* 184–188.

Received December 30, 1977 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 680–700

Human Crowding and Personal Control: An Integration of the Research

Donald E. Schmidt and John P. Keating
University of Washington

Empirical and theoretical discussions have suggested that crowding is experienced when situational density forces the blocking of goals, the interruption of behaviors, or cognitive overload to occur. However, no psychological principles have been employed to unify these explanations. The present article attempts to link the literature on human crowding with the experimental research on personal control. Averill's distinctions among behavioral, cognitive, and decisional control are discussed in the context of human crowding. A conceptual model is offered that suggests that crowding is an attributional label applied to a setting when situational density results in a loss or lack of personal control.

The development of industrialization in the United States has been associated with a shift from a dispersed, agriculturally based society to a relatively centralized urban culture and a period of rapid population growth (Davis, 1965; Meadows, Meadows, Randers, & Behrens, 1972). These general population growth patterns have created dense urban environments in which the majority of the population resides (Hawley, 1971). Perhaps this is why human crowding has become a central topic of research in the emerging subdiscipline of environmental psychology.

Theoretical discussions of crowding have focused on two overriding questions: What physical, social, and personal factors determine an individual's experience of crowding, and what are the psychological and physiological consequences of being crowded on a long-term basis? Although the literature has grown to voluminous proportions, it does not yet offer definitive answers to these questions (Lawrence, 1974).

Initial animal studies established relationships between population density and psychosocial and physiological anomalies. Calhoun (1962) created rat colony densities that were well above levels previously associated with stress. Under these conditions, he found a variety of abnormal developments, including the disturbance of normal maternal behaviors, cannibalism, homosexuality, and the disruption of a number of physiological processes. Christian, Flyger, and Davis (1960) observed the density-related mass mortality of a herd of silka deer on a small island off the Maryland coast. Autopsies of the deer carcasses revealed no consistent evidence of disease or malnutrition, but the presence of enlarged adrenal glands was an apparent sign of stress (cf. Selye, 1956). Results similar to those observed by Calhoun and Christian et al. have also been found in house mice (Southwick, 1955), lemmings (Clough, 1965), and snowshoe hare (Deevey, 1960). Although it is unclear what specific factors contributed to these effects, a plausible explanation has been offered by Christian (1963). He postulated that high levels of population density may have created intraspecies competition for available space and resources. The psychological pressures (stress) that resulted

Preparation of this manuscript was supported in part by University of Washington Graduate School Research Fund Grant RR-07097.

We would like to extend special thanks to Lee Beach, Andrew Davidson, Nickolaus Feimer, and Robert Sasanoff for reviewing initial drafts of this article and to Robert Bolles for his help on references to the animal literature.

Requests for reprints should be sent to Donald Schmidt, who is now at the Societal Analysis Department, General Motors Research Laboratories, Warren, Michigan 48090.

may have culminated in the observed maladies.

Subsequent demographic–correlational studies have attempted to relate these findings to human populations. Associations between population density and indices of crime, disease, and the like commonly have been found (e.g., Galle, Grove, & McPherson, 1972; Hawley, 1972; Schmitt, 1966). However, as provocative as the animal studies and correlational analyses appear, their applicability to conditions encountered in large urban settings is severely restricted. Correlational analyses using relatively gross demographic indices provide only general information about these associations and do not answer questions concerning causation. Similarly, generalization of the results from animal studies to human populations in cities is tenuous. Stages of phylogenetic development have clearly increased the ability of humans to deal with and adapt to conditions in the external environment (Glass & Singer, 1972; Schneirla, 1971; Stokols, 1972b). Thus, the response of other species of animals to external conditions may be noticeably different from that of humankind.

For humans, the relationship between population density and an individual's perception of crowding is far from perfect. Stokols (1972a) has made a useful distinction between these two measures. Population density is defined as a measure of the number of people per unit area, which is strictly a physical index. Alternately, Rapoport (1975) has described functional density as a measure of the number of others within a setting who directly affect an individual's behaviors and perceptions. This extension of the concept specifies a subjective determination of the physical concept. Crowding, on the other hand, is a cognitive evaluation that is predicated on the individual's negative affective reaction to the immediate environment. Stokols (1972a) noted that although density is a necessary antecedent of the perception of crowding, it is not a sufficient cause. Situational constraints, as well as personal and social factors, are also important determinants of this negative evaluation of the environment (Altman, 1975; Carr, 1967; Hall, 1966; Schmidt, in press). Hence, although density must be present for the evaluation to be made, social and dispositional influences also play an important role in the judgment of crowding.

The two primary theoretical approaches developed to explain the crowding response are the behavioral constraint or social interference model and the cognitive overload position. Both approaches incorporate physical, social, and dispositional factors. These explanations have helped to clarify the relationship between density and crowding, since they specify a number of environmental conditions, produced by situational density, that result in an evaluation of crowding. It is important to make clear from the outset of this discussion that some nonspecific level of density is always assumed to be necessary for the production of behavioral constraint, social interference, or cognitive overload. Crowding is expected to result only when the antecedent level of density is present. Second, crowding is treated as the result of the physical consequences of density in these explanations and not as any specific social process resulting merely from the interacting presence of others. A situation is evaluated as crowded when the presence of others results in interference and not when interference is produced only by social interactions with others; that is, crowding has been treated as a physically based and not a socially based reaction to the environment; it is not interference per se but interference related to density that results in an evaluation of crowding.

Behavioral Constraint/Social Interference Explanation

The behavioral constraint or social interference explanation posits that a situation will be evaluated as crowded when density or other related conditions restrict or interfere with the activities of an individual within the setting. This approach is modeled after the theory of psychological reactance, in which Brehm (1966) stated that maintenance of freedom of choice is an important motivating factor in human behavior and perception. He argued that people are disposed to maintain or restore freedom when it

is threatened and that an individual's reaction to a setting is dependent on his or her success at accomplishing this goal. Proshansky, Ittelson, and Rivlin (1970) have discussed reactance in relation to environmental perception. They noted that crowding is experienced when situational density leads to a frustration of an individual's pursuit of important activities and goals. This frustration may be the result of actual physical interference or of the mere presence of others, both of which are construed as limitations of behavioral choices (Stokols, 1972b; Stokols, Rall, Pinner, & Schopler, 1973).

This perspective of limited freedom has been adapted in several theories of crowding. Altman (1975) described social behavior in terms of a privacy-control mechanism that attempts to regulate the frequency of social contacts. According to Altman's model, crowding occurs when the amount of actual social contact exceeds the level that is desired. The person is unable to effectively limit interactions with others. Similarly, Esser (1973) viewed crowding as the result of "not having one's way." This occurs when the denseness of people in a setting creates social interference and hence prevents an individual from functioning effectively. Saegert (1973) noted that high-density environments may create social interference, competition for scarce resources, and the restriction of behaviors that increase perceived crowding. Finally, Stokols (1972b) postulated that an individual experiences crowding when the demand for space required by a specific activity exceeds the available supply.

The research investigating this explanation of crowding has generally provided confirmatory results. Sherrod (1974) conducted an experiment in which three density conditions were manipulated: low density, high density, and high density with control. In the first two conditions, subjects were placed in either dense or nondense settings, but were not given the option of leaving the room. In the third condition, subjects could freely choose to leave the setting. Sherrod found that subjects displayed more decrement in performance on a complex cognitive task and less persistence on an unsolvable puzzle administered subsequent to interaction in a high-density setting relative to a low-density setting. However, perceived control in the high-density setting appeared to ameliorate these negative aftereffects. It is important to note that subjects never actually exercised the option to leave the setting, and hence it was perceived rather than actual control that contributed to these effects.

Sundstrom (1975) manipulated room size and goal blocking independently in an experimental setting. In this study, goal blocking was operationalized as interruption or inattention by others. Subjects reported greater irritation, which increased over time, when goals were blocked than when they were not. However, the amount of self-reported irritation was unrelated to the level of room density. Sundstrom noted that high-density conditions may cause disruption of interpersonal interactions, which subsequently produces psychological stress. On the other hand, density should be unrelated to stress when the number of people present in a setting does not functionally interfere with goal-directed behaviors. Stokols et al. (1973) found that subjects felt more crowded in competitive as compared with cooperative groups, suggesting that the number of people present in a competitive situation may be more salient because they either actively or potentially inhibit the individual from attaining a desired goal or outcome. Hence, the individual is more likely to evaluate the situation in a negative way. Wicker, Kirmeyer, Hanson, Alexander (1976) found that subjects evaluated an experimental setting as more crowded when there were not enough people to perform all of the required tasks, an "undermanned" situation. Perceived crowding was significantly less in the "overmanned" condition, a relatively higher density setting. Since assignments in the undermanned condition required considerable physical movement, often leading to behavioral interference among subjects, this result is supportive of the freedom and control hypothesis. Density is related to crowding when it is expected to reduce the individual's control over the environment. Crowding does not occur when density does not affect behavior or goals.

Schmidt, Goldman, and Feimer (1979) conducted a large-scale field study measuring evaluation of crowding at the residential, neighborhood, and city levels and related these variables to a number of psychological and physical measures. They found that psychological measures indicating some degree of control (e.g., privacy) over the environment were associated with perceived crowding at all levels of the analysis (cf. Altman, 1975; Johnson, 1974). Further, psychological variables became increasingly important and physical measures decreasingly important as one moved from the more immediate residential setting to the less immediate neighborhood and city settings. Schmidt et al. suggested that in the residential setting physical density had a direct effect on the residents' behaviors and activities. However, at the neighborhood and city levels of analysis, crowding evaluations were less tied to objective levels of density and more closely aligned with the impact of more general urban conditions on an individual's behaviors and activities.

Norms, Control, and Crowding

Previous theoretical statements that have been presented concerning the relationship between crowding and perceived freedom and control suggest that normative standards and expectations may have an important effect on evaluations of the environment (Proshansky et al., 1970; Schmidt, in press; Stokols, 1972b; Stokols et al., 1973). Personal control and the subsequent perception of crowding involve individual assessments that are based on the person's expectations. Since norms specify common behavioral standards, they provide predictability that increases control in a social situation (cf. Lefcourt, 1972). While norms increase situational control, violation of such norms decreases control. From this perspective, density leads to an evaluation of crowding when it violates situationally appropriate social or spatial norms. However, the existence of different behavioral and spatial norms among various social subgroups and cultures provides an interesting degree of complexity, a point made especially clear in the anthropological literature (e.g., Hall, 1966; Watsen & Graves, 1966).

Anderson (1972) suggested that high densities may increase the likelihood of highly charged social interactions that cause stress. He noted, however, that in Chinese cultures, large personal spaces are not intrinsically valued and that the existence of well-established behavioral and spatial standards reduces the potential for stressful social interactions induced by close physical proximity, thus reducing the experience of crowding. Draper (1973) reported that among the !Kung bushmen, a hunter–gatherer tribe of southwestern Africa, extremely high-density living is common and preferred. However, periodic escape into outlying desert areas appears to mitigate the potentially stressful effects of high-density living. Finally, Schmidt, Goldman, and Feimer (1976) found that white, black, and Chicano subcultural groups used distinctly different criteria in evaluation of the environment. Crowding was a subjectively different experience for each of the three groups. Black and Chicano subjects tended to focus on a wider range of urban factors in their judgments concerning crowding than did their white counterparts. The above studies suggest that culturally different evaluative and behavioral standards affect environmental judgments.

Several studies have investigated the violation of situationally specific norms as they are related to the evaluation of crowding. Baxter and Deanovich (1970) instructed confederates to position themselves at an inappropriately close or normal spatial distance from an experimental subject. Compared with subjects in the normal condition, subjects in the inappropriately close condition (normative violation) reported that they felt more crowded and displayed higher projective anxiety in describing figures illustrated in a scenario situation in close physical proximity. Freedman, Levy, Buchanan, and Price (1972) tested males and females in small and large experimental rooms. A statistical interaction between sex of the subject and room density was obtained; women felt less crowded in the small (dense) room than did men. Additionally, women

tended to be more cooperative in the small room, whereas men displayed more competitive behaviors. Epstein and Karlin (1975) found similar results in a study of acute crowding. Males tended to form fragmented, competitive groups in the dense condition, whereas females were more cohesive and cooperative. Epstein and Karlin suggested that these effects may be attributable to differential sex norms. Females are expected to share their distress and therefore form more cohesive groups. Males, on the other hand, are expected to hide distress, which purportedly leads to a more fragmented group orientation. Leibman (1969) also has suggested that socialization of sex roles may be associated with different personal-space requirements for men and women. Different criteria may be applied by the sexes when each evaluates a setting and makes a judgment of crowding.

Locus of Control and Crowding

Although most research has been concerned primarily with situational control, certain dispositional characteristics affecting perceived personal control have also been related to the perception of crowding. Research investigating the locus-of-control personality construct (Rotter, 1966) has found differences in reactions to density and in personal-space needs. In an experimental group situation Schopler and Walton (Note 1) found that subjects classified as internals felt less crowded than subjects classified as externals. Similarly, Duke and Nowicki (1972) found that externals had larger personal spaces than internals on a pencil-and-paper measure in which subjects were asked to imagine the approach of a fictitious stranger and to estimate at what distance they would begin to feel uncomfortable. Personal-space distances are thought to act as buffers between the individual and social contacts, hence offering the individual some control over the quality and nature of interpersonal interactions (Hall, 1962, 1966; Sommer, 1969). In the context of the behavioral constraint explanation of crowding, internals presumably have a greater tolerance for interference from the environment, since they perceive a greater degree of personal control over it. Externals, on the other hand, may perceive generally less control over these conditions and consequently may require larger buffer zones. Similarly, less interference from the environment elicits negative evaluations of crowding.

In summary, the theoretical definition of crowding considered above postulates that an environment produces negative affect and is evaluated as crowded when social or physical factors reduce the amount of perceived freedom and control. Increased levels of density create conditions that make it difficult for the individual to function effectively in a setting by blocking goals and interfering with organized behavioral sequences. Since norms set behavioral standards and expectancies that structure social situations, they increase predictability and personal control over a setting. Similarly, normative violations attributable to density should lead to greater evaluations of crowding, since they functionally reduce situational control. However, because normative standards vary between cultural and social groups, different physical and social conditions may lead to violation of different behavioral and spatial standards according to the particular group considered. Finally, perceived locus of control may differentially affect the criteria that the individual applies to a situation in determining whether control can be maintained and if the situation can be labeled as crowded.

Stimulus Overload Explanation

The stimulus overload formulation predicts that a setting is evaluated as crowded when an individual is overwhelmed by the presence of others or when physical conditions in the environment increase the salience of social density. This is an explanation that relies primarily on interference at a perceptual or cognitive level, in contrast with the behavioral restrictions discussed in the previous section. The stimulus overload conceptions of crowding have been based, for the most part, on early urban sociological theories. Attempting to specify the defining qualities of a city, Wirth (1939) noted that urban areas can be characterized in terms

of their size, their density, and the heterogeneity of the encompassed populations. These characteristics often result in the exposure of city residents to excessive levels of physical and social stimulation. Simmel (1950) suggested that urban dwellers conserve psychic energy by developing proportionally fewer and more superficial interpersonal relationships than do individuals residing in rural areas. Presumably this protects a person from becoming overwhelmed by large numbers of social contacts.

An elaboration of the early sociological papers was presented by Milgram (1970). He noted that the many stimuli that impinge on the individual in urban areas may overwhelm the person's cognitive-processing capabilities. Milgram believed that selective attention to only a small subset of the possible social and physical environmental stimuli is a strategy for dealing with this problem. Hence, individuals spend less time with each social contact and develop more superficial interpersonal relationships with others. Wohlwill (1966) suggested that the investigation of the physical environment is largely a study of stimulation, that optimal amounts must be defined and the concept of adaptation level (Helson, 1947) explored. Lee (1966) noted that individuals learn cognitive strategies that help them avoid or modify incoming stimuli from the environment. Such strategies may allow a person to prevent overstimulation and perceptually to filter out the potentially stressful elements of a setting. The privacy-control mechanism postulated by Altman (1975) is also consistent with a stimulus overload interpretation. Crowding results when a level of social stimulation occurs that is greater than that desired by the individual. However, judgments about what is considered to be the optimal level of social contact may vary from situation to situation.

A number of studies have offered support for this position. Desor (1972) asked subjects to place miniature figurines in a model room until they felt that the room was just short of being crowded. The number of entrances to the room and the presence or absence of partitions were varied. She found that more figurines were placed in rooms when there were partitions present and fewer external entrances. Additionally, the type of activity taking place affected judgments about the appropriate number of people in the room. This result is consistent with Altman's model, since partitions and entrances regulate social contacts and different types of activities are associated with different desired levels of social contact. Valins and Baum (1974) found that students tended to rate corridor-style dormitories as more crowded than suite-style dormitories. They suggested that the former design is less efficient at shielding residents from unwanted social stimulation. Baum, Reiss, and O'Hara (1974) found that when a confederate was placed in close proximity to the subjects, individuals were less likely to stop and drink from a water fountain if there was not a screening barrier. Presumably these barriers shielded the fountain user from unwanted visual intrusion or the possible violation of personal space.

Saegert, Mackintosh, and West (1975) found that density was most likely to produce psychological effects under three conditions: (a) when an individual was required to scan and interact in a setting, (b) when density was created by increasing the number of people rather than by varying the room size, and (c) when there were a large number of people present in a setting. This suggests that the salience of social density may be an important part of this effect (cf. Loo, 1973; Rapoport, 1975) and that various activities, social conditions, and tasks focus attention on or away from physical density. Saegert (1973) reported that patrons in a Manhattan department store were less able to recall details of the store in high- versus low-density conditions. This finding is explained by Milgram's (1970) proposal that urban dwellers filter out irrelevant stimuli in the environment to prevent cognitive overload created by a complex and active surrounding.

In summary, stimulus overload explanations of crowding have been based on sociological theories that primarily describe the city in terms of size, density, and the cultural heterogeneity of the encompassed populations. These characteristics combine to

overwhelm the urban dweller's perceptual processing of the environment. Urban residents attempt to evoke strategies that help reduce the occurrence of this cognitive overload. These cognitive and behavioral strategies reduce exposure to the source of the overload or reduce the impact of stimulus overload. Crowding, defined in this way, occurs when the person is unable to cope effectively with the perceptual or cognitive interference created by density. An important component of this model is the notion that optimal levels of stimulation are situationally determined by the individual. The perception of crowding is greatest when the level of stimulation incurred is beyond the level desired and the person is unable to effectively eliminate or reduce such overstimulation. The stimulus overload position is primarily concerned with interference and disruption of perceptual or cognitive information processing.

Crowding and Personal Control Over the Environment

The thesis of the present review is that the theoretical explanations of crowding discussed above all describe either a lack or loss of control over the environment. Density-related conditions functionally reduce the individual's ability to maintain situational control. Second, the perception of crowding is an evaluation in response to this lack or loss of control when density is a salient and viable cause of such absence of control. The behavioral constraint/social interference explanation posits that an environment will be rated as crowded when other people in the setting inhibit or block goals and behaviors. Similarly, the stimulus overload conceptualization posits that the perception of crowding is related to the inability of the individual to control the level of social stimulation.

Personal control is not a simple psychological variable. Rather, it is a complex composite of different concepts that are applicable to the present discussion. These components of control noticeably crosscut the theoretical discussions of crowding that were detailed previously. Averill (1973) has distinguished three primary categories of control: behavioral, cognitive, and decisional.

Behavioral Control

Averill (1973) defined behavioral control in terms of two components: regulated administration and stimulus modifiability. Regulated administration specifies who administers the noxious stimuli (i.e., self-administered by the subject or under external control). Stimulus modifiability refers to how and when the stimulus is encountered. In the experimental research on personal control, stimulus modifiability has involved prevention of noxious stimuli, premature termination of the stimulation, or direct modification of the form of the stimulation. Johnson (1974) similarly defined two components of behavioral control: behavioral selection and outcome effectance. Behavioral selection refers to the presence of viable behavioral options that can be selected by the individual. Johnson noted that the more diverse these options appear, the more personal control is perceived (cf. Steiner, 1970). Outcome effectance refers to the link between initiated behaviors and the final attainment of outcomes or goals. Established relationships between specific behaviors and the attainment of various outcomes increase perceived control by structuring the situation and adding predictability (Lefcourt, 1972). In sum, then, behavioral control involves the ability to choose actions that deal effectively with aversive stimuli and to attain desired outcomes. It also involves coping behaviors that deal with aversive qualities of the environment and aid in the successful pursuit of goals in a specific setting.

In the context of the preceding crowding explanations, behavioral control is applicable in a number of ways. First, the behavioral constraint or social interference position deals with a blockage or inhibition of the behavior–goal relationship. This explanation posits that social density often disrupts behaviors or makes certain outcomes unattainable and that this interference leads to the perception of crowding. Mandler (1964) suggested that the interruption of organized behavioral patterns produces arousal that is

evaluated as positive or negative according to the context of the situation (cf. Schachter & Singer, 1962). Interference that inhibits the relationship between behaviors and outcomes is often disruptive, since it blocks situationally appropriate activities. Hence, this arousal may lead to a negative evaluation of the source of the disruption. When density interferes with goal-directed behaviors, a negative affective attribution of crowding may result (Sundstrom, 1975; Wicker et al., 1976). Similarly, density is not related to negative affect when it does not cause the blocking of goals or the inhibition of behaviors.

Research has provided support for this position. Sundstrom (1975) found that self-reported stress and irritation were not related to the level of density when the blockage of behaviors and goals was independent of density. Wicker et al. (1976) found that there was more perceived crowding in a lower density setting as compared with a higher density setting when behavioral interference occurred in the former but not in the latter. Hence, it is apparent that behavioral interference caused by other people in the setting is a critical factor in determining whether the situation will be evaluated as crowded. Crowding only occurs when antecedent density is a source of the disruption.

The ability to exercise direct behavioral control over an aversive stimulus also reduces negative affective responses. Corah and Boffa (1970) found that control over the termination of a stressor (noise) decreased the subjectively rated aversiveness of the stimulus and reduced physiological indices of arousal (see also, Glass et al., 1973). Similarly, Glass and Singer (1972) found that control over the onset, termination, or predictability of noise reduced negative aftereffects on postexperimental tasks. In this series of studies, subjects perceived that they had control over onset or termination of the stimulus, although no actual control was ever exercised. Other studies have shown lower indices of autonomic arousal for subjects with perceived control over aversive stimuli (e.g., shock) as compared with those without control (e.g., Geer, Davison, & Gatchel, 1970; Mahler, 1973; Starke, 1973). Sherrod (1974) performed a partial replication of the Glass and Singer studies using density as the environmental stressor. He found that compared with the condition of no perceived control, perceived control over the experience of density resulted in a smaller decrement in performance on a postexperimental task. Control in this study involved the freedom to leave a dense setting, although actual control was never exercised. Clearly, perceived control reduces both negative perceptions of a stimulus and aftereffects that may result.

In some situations, however, the individual may be unwilling to assume control when it is possible. Rodin (1976) grouped subjects in a laboratory study on the basis of whether they lived in relatively high- or low-density housing. Subjects were run through a series of experimental tasks in which they either failed or succeeded. She found, in general, that children living in dense housing were less willing to modify the operant task in which they were engaged (i.e., to change reinforcement contingencies by pressing a button). The same children were also more likely to fail on a solvable experimental problem if they had experienced initial failure on an unsolvable problem. Rodin suggested that children living in dense housing may learn that controlling one's environment is difficult. This lack of behavioral control was reaffirmed by outcomes on the initial task.

Coping also provides an alternate method for maintaining behavioral control over an aversive stimulus. When the ability to directly modify the stimulus or to determine when it will be encountered is uncertain, the individual may either implement behaviors that reduce the impact of the stressor or choose to avoid it altogether. The stimulus overload literature suggests that withdrawal is an obvious strategy for coping with excessive stimulation from high-density environments. Milgram (1970) proposed that social withdrawal is a response to cognitive overload created by the presence of large numbers of people. Similarly, the behavioral constraint/social interference explanation of crowding recognizes that situational escape and avoidance of social interactions are ways

of reestablishing control that is threatened in a dense setting (Altman, 1975; Schmidt, in press; Stokols, 1972b).

Research that has looked at coping and crowding has defined withdrawal in one of two ways. The first is in terms of passive avoidance of social interactions. Tucker and Friedman (1972) observed naturally occurring groups at three college locations with surrounding city areas that varied in population density. They found an inverse relationship between group size and city density. They suggested that the establishment of fewer interpersonal contacts may be a strategy for dealing with higher levels of density. Loo (1972) studied groups of young children at play. She noted that as density increased, both the number of social interactions and the incidence of aggressive behavior decreased. Further analyzing these data, she found that while there were no significant differences for girls, boys tended to be more aggressive in the low-density conditions. Hutt and Vaisey (1966) observed that groups of normal children at play interact less in high-density situations and more in low density conditions. Sundstrom (1975) observed less self-disclosure in an experimental setting in which density was high. He observed significantly less eye contact, gesturing, and positive head nodding in manipulated goal-blocking conditions than in conditions in which goal blocking did not occur. In this study the observed behavioral (passive) withdrawal was unrelated to room density. Subjects withdrew from interactions in response to goals that were blocked. Again, Sundstrom manipulated goal blocking and density (room size) as independent factors. Hence, these results offer evidence that the subjective impact of density, rather than objective physical conditions, is important in determining which environmental factors lead to coping.

A second definition of withdrawal involves the active avoidance of social contacts. Altman's (1975) boundary-control model of privacy postulates that the individual may actively seek or thwart social interactions to maintain an appropriate level of social stimulation. Schmidt et al. (1979) found that the ability to attain the desired degree of privacy and the freedom to get away from the residence were important predictors of the self-reported perception of crowding in the residence. These clearly are factors that involve withdrawal or escape from a restricted setting. Kutner (1973) manipulated group size, interpersonal distance, and visual body exposure. He found no relationship between any of these factors and self-rated anxiety; however, behaviors that protected the subject from the visual scrutiny of others increased over time for the high-visual-exposure groups. Kutner suggested that density did not increase anxiety because subjects were able to implement behaviors that eliminated the threat. He postulated that when these behaviors become ineffective in shielding the individual from surveillance by others, anxiety may be increased by the number and proximity of people in the room; that is, density is expected to intensitfy the problem of visual exposure.

A final aspect of behavioral control and coping is anticipatory responding to density. Crowding is expected to occur when the individual is unable to take action to cope with an impending high-density situation following prior warning of the condition. Stokols (1976) noted that anticipation of and preparation for conditions that are potentially stressful are important components of the behavioral constraint explanation. The ability to implement learned responses based on prior experiences with an environmental condition may help mitigate the potentially negative effects prior to the actual onset of a stressor (Appley & Trumbull, 1967; McGrath, 1970). Similarly, Lazarus (1966) has noted that the psychological stress that is experienced by an individual is inversely proportional to the person's anticipated capability to deal with the stressor.

Although the research on anticipatory responses and crowding is sparse, some interesting results have been obtained. Schopler and Walton (Note 1) manipulated anticipated interference by leading subjects to expect tasks that were either structured or unstructured. They reasoned that group structure would act as a regulatory mechanism of interactions between group members

involved in performing a task. Hence, the behavior–outcome relationship (outcome effectance) would be clearly specified. The lack of an established task and group structure was expected to be related to less situational control and a greater potential for interference with other group members, since rules for controlling group interactions and for completing tasks were unspecified. Schopler and Walton found a marginally significant trend: Subjects who believed that the task situation would be unstructured felt more crowded than subjects expecting a structured group. It was suggested that the degree of structure affects anticipated control over the situation.

Baum and Greenberg (1975) conducted a study in which subjects believed that they would be interacting in either 4- or 10-persons groups in experimental rooms of the same size. Although all conditions actually involved groups composed of a subject and two confederates, each subject believed that the entire 4- or 10-person group would meet. It was found that subjects anticipating 10-person groups perceived the room to be more crowded than did those anticipating 4-person groups. Individuals expecting large groups tended to select corner seats in the room, positions from which they could maintain the greatest control over spatial intrusion. Additionally, subjects in the 10-person condition tended to look at the confederates less often and rated each successive confederate entering the room progressively more negatively.

In summary, the evaluation of crowding is related to a loss of behavioral control when physical density creates a condition in which behaviors are interrupted and the attainment of goals is blocked. Research has been presented, however, that indicates that the expectation of control over an aversive stimulus, even in the absence of objective control, is sufficient to forestall negative reactions to the environment. Finally, evidence has been presented that suggests that coping responses initiated prior or subsequent to the onset of a dense condition in which behavioral interference and goal blocking are likely to occur can, in some instances, mitigate negative responses to environmental conditions. Ineffective coping with conditions related to density, on the other hand, is expected to increase evaluations of a setting as crowded.

Cognitive Control

Behavioral control refers to responses that either produce direct action on the environment or allow the individual to actively avoid negative conditions, whereas cognitive control is concerned with the way events or conditions are interpreted by the individual. In this regard, Averill (1973) has defined two elements: information gain and appraisal. Information gain refers to the predictability of a stressor and the anticipation of an aversive event. However, in contrast with the anticipatory responses discussed in the previous section, information gain does not involve any direct action on the environment. Rather, it is concerned with cognitive preparation for an event. Appraisal, the second component, involves the interpretation and evaluation of events. Johnson (1974) defined the term *outcome realization* as the evaluation and interpretation of outcomes, a concept similar to appraisal.

A review of Averill (1973) shows that information concerning the occurrence of an aversive stimulus is related in a complex way to stress. He cited animal studies suggesting that prior warning often increases stress (e.g., overt behavioral disturbances), which can be reduced by using available avoidance responses. Other studies have demonstrated a clear preference for signaled versus unsignaled aversive stimulation, even when escape or avoidance was not possible (e.g., Badia & Culbertson, 1972; Perkins, Seymann, Levis, & Spencer, 1966). Studies with human subjects have also shown a complicated pattern of results. Geer and Maisel (1972) found that subjects who could predict the occurrence of a stressor exhibited greater autonomic responding than did control subjects. Epstein (1973) suggested that although prior warning of an aversive stimulus may initially increase stress, the subsequent generation of accurate expectancies may facilitate habituation, ultimately reducing stress. Monat,

Averill, and Lazarus (1972) found that subjects preferred to know when an electric shock would occur and exhibited less anticipatory responding when this information was not provided. Similarly, Starke (1973) found that the predictable condition led to lower galvanic skin response rates and less subjectively reported discomfort than the unpredictable condition. Staub and Kellett (1972) observed that subjects receiving information concerning the objective and subjective qualities of an aversive stimulus (electric shock) tolerated a more intense level of stimulation before rating it as painful than the level tolerated by subjects provided with only partial or no information. Presumably, this information allows the individual to develop accurate expectancies concerning objective danger and subjective sensation. Hence, information gain functions in two ways. First, it signals the onset of a stressor so that possible cognitive preparation can be initiated. Second, it allows the individual to generate accurate expectancies concerning the stimulus, resulting in eventual habituation to arousal. Lack of appropriate avoidance responses or the generation of inaccurate expectancies may increase the perceived stressfulness of an environmental stimulus (cf. Averill, 1973; Lazarus, 1966).

The importance of information gain is apparent from the literature on human crowding. The role of norms in providing a basis for situational expectancies and predictability was discussed in a previous section. The anticipation of inappropriate spatial violations, the inability to control spatial boundaries, or the blocking of situationally appropriate behaviors may be related to an increase in perceived crowding. Schopler and Walton (Note 1) found a marginally significant trend in which subjects anticipating unstructured groups felt more crowded than groups with well-specified structures. Presumably, unstructured situations increase the potential for social interference. Baum and Greenberg (1975) observed that when subjects anticipated interactions in a dense setting, perceived crowding increased and anticipatory responding was enacted. The behaviors and perceptions observed in these studies were clearly based on the subject's assessment of impending conditions, since expected density levels were never experienced (cf. Lazarus, 1966; Stokols, 1976). Langer and Saegert (1977) reported that subjects felt less uncomfortable and crowded when they were given information about the potential arousal and anxiety they would experience before entering a dense supermarket.

A second effect of information gain is to focus the attention of subjects on the impending occurrence of a stimulus. Averill (1973) has suggested that stimulus overload created by increased vigilance may increase the level of stress experienced prior to the onset of an aversive stimulus. Monat et al. (1972) found that information concerning the time of occurrence of a stressor (indicated by external cues) resulted in increased anticipatory stress for human subjects. Geer and Maisel (1972) also found increased physiological responses to impending stressors by subjects who could predict onset. The results of the Baum and Greenberg (1975) and Schopler and Walton (Note 1) studies can also be interpreted as the consequence of prior warning that effectively focused the subject's attention on the expected qualities of a high-density or potentially uncontrollable situation. This is especially true of the Schopler and Walton study, since subjects in the unstructured condition were initially unaware of what specific conditions they would be facing.

Stokols et al. (1973) have suggested that certain personal and social factors sensitize the individual to actual or potential spatial constraints in a setting. The person may be predisposed to attend to various environmental cues that serve to forewarn him or her of possible restrictions created by social density. Increased sensitivity to the environment is expected to increase the likelihood that a setting will be evaluated as crowded. For example, Schopler and Walton (Note 1) found that subjects with an external locus of control were more likely to evaluate a setting as crowded than were internals. Externals may be more likely to attend to cues that signal a loss of control over the en-

vironment. Schmidt et al. (1976) found that various subcultural groups attended to different factors in their evaluation of crowding. Schmidt (in press) observed that subjects reporting that spatial factors were important considerations in the selection of their current residences were more likely to evaluate the setting as crowded. He postulated that this may reflect an attentional focus on space and spatial restrictions that subsequently occur. Rapoport (1975) suggested that crowding research should focus on density that functionally affects the individual rather than on the gross physical measurements that are typically taken. Clearly, predispositions that focus the individual's attention on the external environment increase this functional density.

Milgram (1970) proposed that experiences in large urban areas may lead to expectations that evoke long-term cognitive strategies for dealing with social overstimulation and the resultant cognitive overload. Environmental cues that forewarn the individual of potential stimulus overload from social sources may elicit these strategies. For example, Saegert's (1973) department store study (cited previously) offers support for this idea; subjects attended to less detail in the setting. Baum and Greenberg (1975) found that subjects anticipating a dense situation tended to look at others in the setting less often than those anticipating a relatively low-density setting. Sundstrom (1975) also found that goal blocking, a consequence of density in other research, led to less attention by subjects to others present in the room. These behaviors allow the individual to tune out potentially stressful aspects of the setting.

The human crowding literature that has been reviewed demonstrates that information gain given as prior warning of an impending dense situation serves to increase the likelihood that a setting will be evaluated as crowded. The research has suggested that habituation to the arousal resulting from the expectation of high density may be caused by the enactment of coping strategies that perceptually reduce the perceived impact of density by focusing the individual's attention away from the immediate environment. Conditions or predispositions that focus the individual's attention on the implications of density are expected to increase the likelihood of crowding evaluations (cf. Averill, 1973; Baum & Greenberg, 1975; Schmidt, in press). Additionally, anticipation of a dense situation was shown to lead to avoidance behaviors that demonstrated an active attempt to seek protection from anticipated problems. These behaviors suggest that much of the contradictory experimental evidence relating predictability, autonomic arousal, and subjective stress can be reconciled by considering information gain in conjunction with the second component of cognitive control, appraisal.

Appraisal refers to the way that individuals evaluate and interpret events and is related to the subsequent affective responses that can be expected. This aspect of control is important, since crowding is typically defined as a negative affective reaction to conditions created by density. When actual or potential goal blocking can be attributed to the density of a setting, one would expect the individual to evaluate the situation as crowded (Baum & Greenberg, 1975; Stokols et al., 1973; Sundstrom, 1975). Lazarus (1966) noted that stress occurs when the individual anticipates an inability to cope with an impending situation. If the person judges that interference or overload created by density cannot be effectively dealt with, the perception of crowding is increased (cf. Kutner, 1973). Again, one is dealing with a cognitive evaluation that is separate from any behavioral responses.

Stokols (1976) noted that the critical factor in psychological stress is the individual's belief that he or she is unable to exert control over a situation. Crowding implies that the individual perceives an inability to deal with conditions created by density and hence anticipates or experiences goal blockage, spatial restrictions, or cognitive overload. Cognitive or behavioral coping responses are ways of preparing for and dealing with these conditions. Appraisal may, therefore, involve a series of environmental assessments, responses, and reassessments (cf. Stokols,

1972b, 1976). However, cultural and sex-specific norms have been shown to shape these assessments; that is, different behavioral and spatial standards are used to appraise impending or ongoing situations (Anderson, 1972; Draper, 1973; Epstein & Karlin, 1975; Freedman et al., 1972; Hall, 1966; Schmidt et al., 1976; Watsen & Graves, 1966).

A number of studies have demonstrated the effect of appraisal on environmental evaluations. Schopler and Walton (Note 1) found that subjects expecting to interact in unstructured groups reported more perceived crowding than subjects expecting a well-specified group structure. They suggested that this difference may be attributable to the subject's appraisal that the unstructured situation would be less controllable. Sherrod (1974) concluded that the stress of interacting in a high-density situation was reduced by simply making subjects believe that they were free to leave. Desor (1972) found that entrances and partitions in a model room were viewed as ways of maintaining control over interactions (cf. Altman, 1975) and affected the perception of crowding. Partitions and entrances should lead to the appraisal that a setting is controllable.

In summary, appraisal is a cognitive evaluation that is affected by the individual's anticipation of his or her ability to deal with an impending situation. The perception of ability to deal with an impending noxious stimulus can reduced both stress and objective ratings of the aversiveness of the negative affective response to a setting when the individual believes that density-related conditions will create an uncontrollable situation.

Decisional Control

The final element, decisional control, is defined by both Averill (1973) and Johnson (1974) as choice in the selection of specific outcomes or goals. It is the primary component of control in Brehm's (1966) theory of psychological reactance. Our discussion of this element of control is brief, since in the literature on crowding, a limitation of choice or freedom is tantamount to a loss of both behavioral and cognitive control.

From the perspective of the behavioral constraint explanation, crowding is believed to occur when the individual is unable to attain desired outcomes or goals in a setting because density has functionally blocked the behavior–goal relationship. In this way, both behaviors and goals are restricted. Goal blocking and behavioral interference have been related to an increased perception of crowding when density is a viable cause (Proshansky et al., 1970; Saegert, 1973; Schmidt, in press; Schmidt et al., 1979; Wicker et al., 1976). These restrictions functionally limit the options available to an individual in a setting, reducing decisional control and increasing negative affect and psychological reactance.

Brehm (1966) and Wortman and Brehm (1975) noted that the importance of the outcomes that were restricted was a central factor in determining the magnitude of the reactance that was evoked. The blocking of important or primary behavioral alternatives or goal options is expected to produce more reactance than options that are of little personal importance. A similar logic has been proposed by Stokols (1976) in relation to the evaluation of crowding. He suggested that perceived crowding tends to be of greater intensity in primary versus secondary environments. Primary environments include those settings in which the individual spends a great deal of time (e.g., work or residential areas), is engaged in relatively long-term interactions with others, and is involved in a number of personally important activities. Secondary environments are locations in which interactions with others are relatively transitory, anonymous, and of generally less consequence (e.g., public transportation, shopping facilities, etc.). Stokols suggested that goal blocking or behavioral interference in primary environments may involve the limitations of important goals and activities. Hence, the impact of dense or related conditions is more critical, and the perception of crowding should be of a greater magnitude.

Altman's (1975) boundary-control model

of privacy suggests that a loss of decisional control is related to the evaluation of crowding. This theoretical perspective posits that the individual in a social situation decides on an optimal level of social interaction or stimulation. This desired level is either attained or blocked, depending on conditions in the setting (e.g., density, spatial restriction, etc.). If situational conditions force a level of social interaction or stimulation beyond that desired by the individual, an evaluation of crowding is expected to result.

Decisional control refers to the ability of the individual to select outcomes or goals in a social situation. However, in the crowding literature, decisional control is tantamount to behavioral control, since restricted behaviors and blocked goals functionally limit the activities that are possible in an environment. Similarly, an appraisal that a setting is uncontrollable or prior warning of impending conditions may also indicate a reduction in decisional control. The increased likelihood that a setting will be evaluated as crowded is expected to occur if density is a viable cause of these restrictions of behavior and goal alternatives.

Personal Control and the Attribution of Crowding

Although the literature on personal control and the literature on human crowding fit together reasonably well, it is clear that the evaluation of crowding within a specific setting may be the result of deficiencies in a number of control components. The results from a number of the crowding studies are explicable by using one or more of these elements. Certainly the consequence of living and interacting in high-density environments is likely to be a loss of some degree of control over behaviors and the pursuit of goals. It is important at this point to examine the specific cognitive processes that are involved when an individual experiences or anticipates a loss of control and evaluates a setting as crowded.

The literature reviewed previously indicates that a lack of behavioral control or the appraisal that a situation is uncontrollable is related to increased levels of self-reported stress and indices of physiological arousal (e.g., Corah & Boffa, 1970; Geer et al., 1970; Glass & Singer, 1972; Staub, Tursky, & Schwartz, 1971). Similarly, the human crowding literature has suggested that environmental conditions that reduce an individual's control (i.e., goal blocking, behavioral interference, and stimulus overload) also lead to more self-reported stress and arousal (e.g., Saegert, 1973, 1975; Stokols et al., 1973; Sundstrom, 1975; Valins & Baum, 1974). A subsequent evaluation of the environment as crowded is expected to result when density levels are sufficiently high; this evaluation may be tied to loss of control.

A number of studies have demonstrated an association between density and arousal, as measured by skin conductance (Aiello, Epstein, & Karlin, 1975), heart rate and blood pressure (D'Atri, 1975; Evans, 1975), and palmar sweat (Saegert, 1975). However, it is unlikely that the relationship between density and arousal is as simple as these results suggest. In these studies, conditions indicative of some social impact are often confounded with density. It is unclear whether density causes arousal or whether arousal is due to a concomitant consequence. Certainly the behavioral constraint and cognitive overload explanations suggest the latter possibility. Sundstrom (1975) observed that self-reported stress and irritation were not related to density when goal blocking and behavioral interference were manipulated independently of this density. These latter conditions were also related to increased arousal in previous studies (e.g., Mandler, 1964; Wortman & Brehm, 1975).

Findings such as these suggest that physical density may have an indirect effect on arousal. Previous articles have also indicated that anticipated problems precipitated by density may lead to the appraisal that a situation is uncontrollable, thus creating stress and arousal. However, even in this case anticipated goal blocking and behavioral interference are the critical factors resulting in the perception of crowding.

Arousal and stress created by a loss of control over the environment are similar to

what Brehm (1966) has called psychological reactance. Goal blocking, behavioral interference, and stimulus overload are often density-related consequences that decrease an individual's situational control and hence are expected to elicit some degree of reactance. We predict a lower level of arousal and stress associated with a relatively controllable setting and greater arousal and negative affect related to uncontrollable situations. Additionally, we expect the actual or anticipated ability of the individual to cope with or to eliminate the source that threatens personal control to have a direct relationship to the magnitude and duration of the reactance that is experienced (cf. Brehm, 1966). Although reactance theory posits effects that may be attributed to a loss of control over the environment, we expect similar effects for situations in which one finds a general lack of control as well. Presumably, in these settings a lack of control manifests itself in the inability to deal with a situation or to pursue activities and goals. Hence, lack of control in a setting should produce the same arousal and stress effects that have been found in situations in which conditions have precipitated a loss of environmental control. The previous discussions on cognitive control suggest that the individual may manifest initial stress if lack of control is expected and the situation is thus appraised in this way. For the sake of brevity in this discussion, we assume that both a lack and a loss of control are similar conditions; that is, both create increased levels of arousal and can be produced by situational density.

Social psychological theory and research suggest that the link between arousal and the evaluation of a setting as crowded may involve a relatively complex cognitive process. Current research indicates that when arousal is experienced by the individual, it is subsequently labeled as consistent with the context of the situation. Hence, the individual must scan the external environment and attribute arousal to a salient and viable source that is present. Schachter and Singer (1962) conducted a critical investigation of this process. In their initial study, they aroused subjects by injecting them with the drug epinephrine. Subjects were then informed that the drug would cause them to feel aroused (informed condition), would have effects other than arousal (misinformed condition), or would have no noticeable effects (ignorant condition). Schachter and Singer found that subjects were more likely to report emotional states that were consistent with external environmental cues (euphoria or anger) in the misinformed and ignorant conditions than in the informed and placebo (no drug) conditions. They suggested that in the conditions in which subjects did not anticipate arousal, it was necessary for subjects to locate a source in the environment that could be used to cognitively label the internal state. In the informed and placebo conditions, labels were either provided by the experimenter or were unneeded when no physiological arousal was experienced.

While the Schachter and Singer results deal with the cognitive labeling of drug-induced arousal, similar results have been found for situationally induced arousal. These latter studies are more important for the present discussion of crowding, since arousal associated with density and crowding is presumably induced by events occurring in the external environment. The studies that best illustrate the attribution of situationally induced arousal are best described as *misattribution* studies. In these investigations, arousal is induced by making subjects expect the impending occurrence of an environmental stressor (e.g., electric shock). Subjects are led to believe that the real cause of arousal is something else in the experimental setting. Behavioral or self-report measures are then taken to indicate how arousal is labeled. Valins and Ray (1967) found that they could reduce phobias about snakes by convincing subjects that their arousal was due to impending electrical shocks rather than to snakes. Storms and Nisbett (1970) found that they could reduce insomnia by making subjects believe that their arousal was due to a pill (placebo) administered to them rather than to their personal problems. Finally, Ross, Rodin, and Zimbardo (1969) found

that subjects were more willing to work for monetary rewards than to escape from impending electrical shocks when they were convinced that their arousal was due to the presence of a loud noise rather than to the shocks.

The idea that the evaluation of crowding may be an attributional response to arousal has been a relatively recent development in the literature. A number of studies have offered support for this notion. Freedman (1975) provided evidence suggesting that density acts to intensify positive or negative feelings according to the context of the situation. The environment provides information concerning the interpretation of arousal, and density effectively heightens the level of positive or negative affect. Gochman (1977; see also Keating, 1978; Kalb & Keating, Note 2) conducted a series of studies that directly investigated this labeling process. She found that unfulfilled goals or disconfirmed expectations were related to an evaluation of the setting as crowded when density was salient. She suggested that disconfirmed expectations and unmet goals create a state of arousal (cf. Lazarus, 1966; Mandler, 1964), which is subsequently attributed or misattributed to the most salient environmental condition.

The attributional explanation posits that arousal created by a loss of control may lead to an evaluation of crowding if social factors create some sort of behavioral or perceptual interference and situational density is a salient environmental factor. An interesting test of this approach was provided by Worchel and Teddlie (1976). In their study room size (spatial density; Loo, 1973), close or far interaction distance, and the presence or absence of picture distractors on the walls were manipulated as independent variables. They found that the perception of crowding was greatest when density was high, when interaction distance was close, and when picture distractors were not present. When interaction distance was close and distractions were present, the room was rated as relatively less crowded. Worchel and Teddlie suggested that close interaction distances caused spatial violations that were associated with increased levels of arousal (Felipe & Sommer, 1966; Hall, 1966; Sommer, 1969). When the subject searched the setting for a cause of the arousal, attention was drawn to either density or to the visual distractors. Hence, when attention was focused away from room density by the picture distractors, arousal was not attributed to density and an evaluation of crowding was less likely. However, when density was salient, the setting was evaluated as significantly more crowded. Langer and Saegert (1977) found that subjects performing an experimental task in a dense supermarket reported feeling less crowded and more comfortable when they were told that they might become aroused and anxious than when they were uninformed. Hence, information gain concerning potential internal states may partially alleviate negative affect attributable to dense settings.

Research investigating the attribution approach to crowding has suggested that the evaluation of crowding may either be a veridical or a nonveridical response to the environment; that is, density may create conditions that lead to increased arousal and a subsequent evaluation of crowding, or arousal may be misattributed to density when it is merely a salient, although noncausative, factor in the setting (Keating, 1978). In the present discussion, we are most interested in the former case (when crowding is a veridical response based on the impact of density), since this has been the primary theoretical and empirical focus in the past.

Control–Attribution Model of Human Crowding

Figure 1 is an illustration of the suggested process. Previous research and theory has described crowding in terms that imply a loss of control over the environment. Although absolute density has been shown to have an inconsistent relationship to the cognitive evaluation of crowding, functional density in its impact on the individual's activities and behaviors appears to be important (Rapoport, 1975). Ample evidence has been presented that indicates that loss of control is associated with physiological

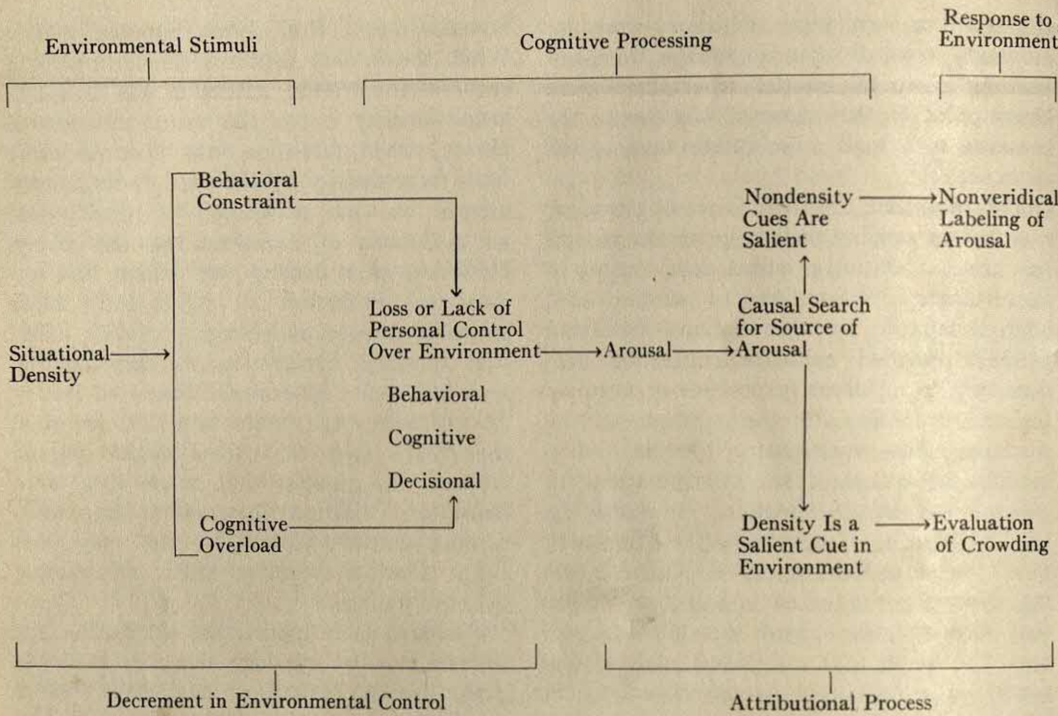

Figure 1. A control-attribution model of human crowding.

and self-reported arousal, especially when the situation creates a decrement in behavioral control or the appraisal that a situation is uncontrollable. Similarly, goal blocking, behavioral interference, and stimulus overload have been viewed as sources of arousal or stress in previous articles. This arousal state that is produced by a lack of or loss of control over the environment is a necessary antecedent to the subsequent attributional process.

The attribution segment of the diagram details the cognitive events that are thought to take place when arousal is experienced and a cognitive label is needed. The individual scans the environment in an attempt to find a viable source of the internal physiological state. However, increasing levels of arousal lead to a restriction of the number of environmental cues that are considered and of the amount of information incorporated in the individual's decisions and judgments (Easterbrook, 1959; Schroder, Driver, & Streufert, 1967). Hence, the salience of various environmental cues becomes important. Specifically, cues that are salient and are perceived as viable causes of the arousal state are selected because of the limited attentional focus maintained by the individual (cf. Worchel & Teddlie, 1976). These cues may then provide the cognitive label for the arousal that is experienced. In the present case, crowding is viewed as an attributional label when density is salient and has acted to functionally limit an individual's personal control over a setting. This model clearly allows for either veridical or nonveridical labeling of arousal generated by density-related conditions, although the former condition is our primary interest in this article.

The present model makes some distinctions that are relevant to follow-up research in the area of human crowding. Figure 1 suggests three distinct stages in the evaluation of crowding: the presence of environmental stimuli and objective environmental conditions, cognitive processing requiring an assessment of personal control and the subsequent attributional phase, and the final evaluation of crowding. Although it may seem tenuous to classify behavioral con-

straint or cognitive overload as components of environmental stimuli, this is precisely the way these concepts have been operationalized in the experimental research; they are objectively manipulated conditions. In this sense, previous views of crowding have involved a simple stimulus–response explanation for crowding. Objective environmental events or stimuli lead to the evaluation of crowding. The control–attribution approach includes these previous explanations along with indications of interceding cognitive processes. It is clear that an investigation of the nature of these processes, especially of the components of personal control, is a necessary step for a full understanding of crowding. It seems important to map conditions such as behavioral constraint and cognitive overload onto subjective control and to investigate individual-difference variables (e.g., demographic variables, stable personality dispositions, etc.) as mediators of this relationship. With this type of research it should be possible to make more accurate predictions concerning the evaluation of crowding than would be allowed by a simple stimulus–response model and to tie together both physical environmental stimuli and subsequent cognitive evaluations.

Social psychological theory and the experimental research on personal control suggest that the literature on human crowding is explicable by using established psychological principles. The application of these more general phenomena offers both a consolidation of diverse explanations of crowding and a logical thread tied to larger empirical bases. However, a clear need for more detailed study of the impact of physical and social conditions on behaviors and on subsequent evaluations such as crowding has been established. It is these additional studies that will help to map the relationships among objective environmental conditions and the psychological effects they produce.

Reference Notes

1. Schopler, J., & Walton, M. *The effects of expected structure, expected enjoyment, and participant's internality-externality upon feelings of being crowded.* Unpublished manuscript, University of North Carolina at Chapel Hill, 1974.
2. Kalb, L., & Keating, J. P. *Non-spatial factors affecting crowding attributions in a dense field setting.* Unpublished manuscript, University of Washington, 1978.

References

Aiello, J. R., Epstein, Y. M., & Karlin, R. A. Effects of crowding on electrodermal activity. *Sociological Symposium*, 1975, *14*, 43–57.

Altman, I. *The environment and social behavior.* Monterey, Calif.: Brooks/Cole, 1975.

Anderson, E. N. Some Chinese methods of dealing with crowding. *Urban Anthropology*, 1972, *1*, 141–150.

Appley, M. H., & Trumbull, R. On the concept of psychological stress. In M. H. Appley & R. Trumbull (Eds.), *Psychological stress.* New York: Appleton-Century-Crofts, 1967.

Averill, J. R. Personal control over aversive stimuli and its relationship to stress. *Psychological Bulletin*, 1973, *80*, 286–303.

Badia, P., & Culbertson, S. The relative aversiveness of signalled vs. unsignalled escapable and inescapable shock. *Journal of the Experimental Analysis of Behavior*, 1972, *17*, 463–471.

Baum, A., & Greenberg, C. I. Waiting for a crowd. *Journal of Personality and Social Psychology*, 1975, *32*, 671–679.

Baum, A., Reiss, M., & O'Hara, J. Architectural variants of reaction to spatial invasion. *Environment and Behavior*, 1974, *6*, 91–100.

Baxter, J. C., & Deanovich, B. F. Anxiety arousing effects of inappropriate crowding. *Journal of Consulting and Clinical Psychology*, 1970, *35*, 174–178.

Brehm, J. *A theory of psychological reactance.* New York: Academic Press, 1966.

Calhoun, J. Population density and social pathology. *Scientific American*, 1962, *206*(2), 139–150.

Carr, S. The city of the mind. In J. W. Ewald, Jr. (Ed.), *Environment for man: The next fifty years.* Bloomington: Indiana University Press, 1967.

Christian, J. Endocrine adaptive mechanisms and the physiologic regulation of population growth. In W. Mayer & R. Gelder (Eds.), *Physiological mammalogy* (Vol. 1). New York: Academic Press, 1963.

Christian, J., Flyger, V., & Davis, D. Factors in the mass mortality of a herd of silka deer (*Cervus nippon*). *Chesapeake Science*, 1960, *1*, 79–95.

Clough, G. C. Lemmings and population problems. *American Scientist*, 1965, *53*, 199–212.

Corah, N. L., & Boffa, J. Perceived control, self-observation, and response to aversive stimulation. *Journal of Personality and Social Psychology*, 1970, *16*, 1–4.

D'Atri, D. A. Psychophysiological responses to crowding. *Environment and Behavior*, 1975, *7*, 237–252.

Davis, K. The urbanization of the human population. *Scientific American,* 1965, *213*(3), 40–53.
Deevey, F. S. The hare and the haruspex. *American Scientist,* 1960, *48,* 415–429.
Desor, J. A. Toward a psychological theory of crowding. *Journal of Personality and Social Psychology,* 1972, *21,* 79–83.
Draper, P. Crowding among hunter-gatherers: The !Kung bushmen. *Science,* 1973, *182,* 301–303.
Duke, M., & Nowicki, S. A new measure and social learning model for interpersonal distance. *Journal of Experimental Research in Personality,* 1972, *6,* 119–132.
Easterbrook, J. A. The effects of emotion on cue utilization and the organization of behavior. *Psychological Review,* 1959, *66,* 183–201.
Epstein, S. Expectancy and magnitude of reaction to a noxious UCS. *Psychophysiology,* 1973, *10,* 100–107.
Epstein, Y., & Karlin, R. A. Effects of acute experimental crowding. *Journal of Applied Social Psychology,* 1975, *5,* 34–53.
Esser, A. H. Experiences of crowding: Illustration of a paradigm for man-environment relations. *Representative Research in Social Psychology,* 1973, *4,* 207–218.
Evans, G. W. Behavioral and physiological consequences of crowding in humans (Doctoral dissertation, University of Massachusetts—Amherst, 1975). *Dissertation Abstracts International,* 1975, *36,* 4725B. (University Microfilms No. 76-5277)
Felipe, N. J., & Sommer, R. Invasions of personal space. *Social Problems,* 1966, *14,* 206–214.
Freedman, J. L. *Crowding and behavior.* San Francisco: Freeman, 1975.
Freedman, J. L., Levy, A., Buchanan, R., & Price, J. Crowding and human aggression. *Journal of Experimental Social Psychology,* 1972, *8,* 528–548.
Galle, O., Grove, W., & McPherson, J. Population density and pathology, *Science,* 1972, *176,* 23–30.
Geer, J. H., Davison, G. C., & Gatchel, R. I. Reduction of stress in humans through nonveridical perceived control over aversive stimulation. *Journal of Personality and Social Psychology,* 1970, *16,* 731–738.
Geer, J. H., & Maisel, E. Evaluating the effects of the prediction-control confound. *Journal of Personality and Social Psychology,* 1972, *23,* 314–319.
Glass, D. C., & Singer, J. E. *Urban stress.* New York: Academic Press, 1972.
Glass, D. C., et al. Perceived control of aversive stimulation and the reduction of stress responses. *Journal of Personality,* 1973, *41,* 577–595.
Gochman, I. R. Causes of perceived crowding unrelated to density. (Doctoral dissertation, University of Washington, 1976). *Dissertation Abstracts International,* 1977, *37,* 3675B. (University Microfilms No. 77-576)
Hall, E. T. The madding crowd: Space and its organization as a factor in mental health. *Landscape,* 1962, *11,* 26–29.
Hall, E. T. *The hidden dimension.* New York: Doubleday, 1966.
Hawley, A. H. *Urban society.* New York. Ronald Press, 1971.
Hawley, A. H. Population density and the city. *Demography,* 1972, *9,* 521–529.
Helson, H. Adaptation-level as frame of reference for prediction of psychophysical data. *American Journal of Psychology,* 1947, *60,* 1–29.
Hutt, C., & Vaisey, M. Differential effects of group density on social behavior. *Nature,* 1966, *209,* 1371–1372.
Johnson, C. A. Privacy as personal control. In W. Preiser (Ed.), *Proceedings of the Environmental Design Research Association.* Stroudberg, Pa.: Dowden, Hutchinson & Ross, 1974.
Keating, J. P. Scapegoating the environment: Toward a model of crowding. In, *Priorities for Environmental Design Research.* Washington, D.C.: Environmental Design Research Association, 1978.
Kutner, D. H., Jr. Overcrowding: Human response to density and visual exposure. *Human Relations,* 1973, *26,* 31–50.
Langer, E. J., & Saegert, S. Crowding and cognitive control. *Journal of Personality and Social Psychology,* 1977, *35,* 175–182.
Lawrence, E. S. Science and sentiment: Overview of research on crowding and human behavior. *Psychological Bulletin,* 1974, *81,* 712–720.
Lazarus, R. *Psychological stress and the coping process.* New York: McGraw-Hill, 1966.
Lee, D. H. K. The role of attitude in response to environmental stress. *Journal of Social Issues,* 1966, *12,* 83–91.
Lefcourt, H. M. Recent developments in the study of locus of control. In B. A. Maher (Ed.), *Progress in experimental personality research* (Vol. 6). New York: Academic Press, 1972.
Leibman, M. The effects of sex and race norms on personal space. *Environment and Behavior,* 1969, *2,* 208–246.
Loo, C. The effects of spatial density on the social behavior of children. *Journal of Applied Social Psychology,* 1972, *2,* 372–381.
Loo, C. Important issues in researching the effects of crowding on humans. *Representative Research in Social Psychology,* 1973, *2,* 327–381.
Mahler, C. R. The effects of perceived gain, loss, and absence of control on stress behavior during threat (Doctoral dissertation, State University of New York at Buffalo, 1972). *Dissertation Abstracts International,* 1973, *33,* 3951B. (University Microfilms No. 73-5144)
Mandler, G. The interruption of behavior. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 12). Lincoln: University of Nebraska Press, 1964.
McGrath, J. E. Major substantive issues: Time, setting, and the coping process. In J. E. McGrath (Ed.), *Social and psychological factors in stress.* New York: Holt, Rinehart & Winston, 1970.

Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W., III. *The limits to growth.* Washington, D.C.: Potomac Books, 1972.

Milgram, S. The experience of living in cities. *Science,* 1970, *167,* 1461–1468.

Monat, A., Averill, J. R., & Lazarus, R. S. Anticipatory stress and coping reactions under various conditions of uncertainty. *Journal of Personality and Social Psychology,* 1972, *24,* 237–253.

Perkins, C. C., Jr., Seymann, R. G., Levis, D. J., & Spencer, H. R., Jr. Factors affecting preference for signal-shock over shock-signal. *Journal of Experimental Psychology,* 1966, *72,* 190–196.

Proshansky, H. M., Ittelson, W. H., & Rivlin, L. G. Freedom of choice and behavior in a physical setting. In H. M. Proshansky, W. H. Ittelson, & L. G. Rivlin (Eds.), *Environmental psychology.* New York: Holt, Rinehart & Winston, 1970.

Rapoport, A. Towards a redefinition of density. *Environment and Behavior,* 1975, *7,* 133–158.

Rodin, J. Density, perceived choice and response to controlled and uncontrolled outcomes. *Journal of Experimental Social Psychology,* 1976, *12,* 564–578.

Ross, L., Rodin, J., & Zimbardo, P. G. Toward an attribution therapy: The reduction of fear through induced cognitive–emotional misattribution. *Journal of Personality and Social Psychology,* 1969, *12,* 279–288.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs,* 1966, *80*(1, Whole No. 609).

Saegert, S. Crowding: Cognitive overload and behavioral constraint. In W. Preiser (Ed.), *Proceedings of the Environmental Design Research Association* (Vol. 2). Stroudberg, Pa.: Dowden, Hutchinson & Ross, 1973.

Saegert, S. C. Effects of spatial and social density on arousal, mood, and social orientation (Doctoral dissertation, University of Michigan, Ann Arbor, 1974). *Dissertation Abstracts International,* 1975, *35,* 3649B. (University Microfilms No. 75-793)

Saegert, S., Mackintosh, E., & West, S. Two studies of crowding in urban public spaces. *Environment and Behavior,* 1975, *7,* 159–184.

Schachter, S., & Singer, J. E. Cognitive, social, and physiological determinants of emotional state. *Psychological Review,* 1962, *69,* 379–399.

Schmidt, D. E. Crowding in urban environments: An integration of theory and research. In J. Aiello (Ed.), *Residential crowding and design.* New York: Plenum Press, in press.

Schmidt, D. E., Goldman, R. D., & Feimer, N. R. Physical and psychological factors associated with perceptions of crowding: An analysis of subcultural differences. *Journal of Applied Psychology,* 1976, *61,* 279–289.

Schmidt, D. E., Goldman, R. D., & Feimer, N. R. Perceptions of crowding: Predicting at the residence, neighborhood and city levels. *Environment and Behavior,* 1979, *11,* 105–130.

Schmitt, R. C. Density, health, and social disorganization. *Journal of the American Institute of Planners,* 1966, *32,* 38–40.

Schneirla, T. C. The concept of development in comparative psychology. In J. Eliot (Ed.), *Human development and cognitive processes.* New York: Holt, Rinehart & Winston, 1971.

Schroder, H. M., Driver, M. J., & Streufert, S. *Human information processing.* New York: Holt, Rinehart & Winston, 1967.

Selye, H. *The stress of life.* New York: McGraw-Hill, 1956.

Sherrod, D. R. Crowding, perceived control, and behavioral aftereffects. *Journal of Applied Social Psychology,* 1974, *4,* 171–186.

Simmel, G. The metropolis and mental life. In K. Wolff (Ed.), *The sociology of George Simmel.* New York: Free Press, 1950.

Sommer, R. *Personal space: The behavioral basis of design.* Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Southwick, C. H. The population dynamics of confined house mice supplied with unlimited food. *Ecology,* 1955, *36,* 212–225.

Starke, M. C. Effects of control and prediction on reactions to aversive stimulation (Doctoral dissertation, State University of New York at Stony Brook, 1972). *Dissertation Abstracts International,* 1973, *33,* 5551B. (University Microfilms No. 73-10,903)

Staub, E., & Kellett, D. S. Increasing pain tolerance by information about aversive stimuli. *Journal of Personality and Social Psychology,* 1972, *21,* 198–203.

Staub, E., Tursky, B., & Schwartz, G. E. Self-control and predictability: Their effects on reactions to aversive stimulation. *Journal of Personality and Social Psychology,* 1971, *18,* 157–162.

Steiner, I. D. Perceived freedom. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 5). New York: Academic Press, 1970.

Stokols, D. On the distinction between density and crowding. *Psychological Review,* 1972, *79,* 275–277. (a)

Stokols, D. A social-psychological model of human crowding phenomena. *Journal of the American Institute of Planners,* 1972, *38,* 72–82. (b)

Stokols, D. The experience of crowding in primary and secondary environments. *Environment and Behavior,* 1976, *8,* 49–86.

Stokols, D., Rall, M., Pinner, B., & Schopler, J. Physical, social and personal determinants of the perception of crowding. *Environment and Behavior,* 1973, *5,* 87–115.

Storms, M., & Nisbett, R. E. Insomnia and the attribution process. *Journal of Personality and Social Psychology,* 1970, *2,* 319–328.

Sundstrom, E. An experimental study of crowding: Effects of room size, intrusion, and goal blocking

on nonverbal behaviors, self-disclosure, and self-report. *Journal of Personality and Social Psychology*, 1975, *32*, 645–654.

Tucker, J., & Friedman, S. T. Population density and group size. *American Journal of Sociology*, 1972, *77*, 742–749.

Valins, S., & Baum, A. Residential group size, social interaction, and crowding. *Environment and Behavior*, 1974, *5*, 421–439.

Valins, S., & Ray, A. Effects of cognitive desensitization on avoidance behavicr. *Journal of Personality and Social Psychology*, 1967, *7*, 345–350.

Watsen, O. M., & Graves, T. C. Quantitative research in proxemic behavior. *American Anthropologist*, 1966, *68*, 971–985.

Wicker, A. W., Kirmeyer, S. L., Hanson, L., & Alexander, D. Effects of manning level on subjective experiences, performance, and verbal interaction in grcups. *Organizational Behavior and Human Performance*, 1976, *17*, 251–274.

Wirth, L. Urbanism as a way of life. *American Journal of Sociology*, 1939, *44*, 1–24.

Wohlwill, J. F. The physical environment: A problem for a psychology of simulation. *Journal of Social Issues*, 1966, *12*, 29–38.

Worchel, S., & Teddlie, C. The experience of crowding: A two-factor theory. *Journal of Personality and Social Psychology*, 1976, *34*, 30–40.

Wortman, C. B., & Brehm, J. W. Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 8). New York: Academic Press, 1975.

Received February 17, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 701-720

Therapeutic Videotape and Film Modeling: A Review

Mark H. Thelen, Richard A. Fry,
Peter A. Fehrenbach, and Nanette M. Frautschi
University of Missouri—Columbia

This article reviews research involving the use of videotape and film modeling in clinical and analogue settings. Most of the research has been on phobias, test anxiety, dental and medical stress, and interpersonal skills. Critical methodological issues and the relevance of this research to modeling theory are also discussed. The need for improved methodology and the application of videotape and film modeling in more clinical populations is emphasized.

Symbolic modeling in the form of films or videotapes was the logical outgrowth of the documented effects of live modeling in therapy (Bandura, 1969; Geer & Turtletaub, 1967; Sarason, 1968). It is likely that the cognitive processes identified as important in the effects of live modeling—such as verbal and imaginal coding and rehearsal (Bandura, 1977; Bandura, Blanchard, & Ritter, 1969)—can as well occur as a function of observing a filmed model. There is little reason to believe that the symbolic processes are particularly different as a function of observing a live versus a filmed model (Bandura & Barab, 1973).

Film and videotape modeling has several advantages over live modeling. The availability of filmed and videotaped media provides the opportunity to capture naturalistic modeling sequences that would be difficult to create in clinic settings. And of course the therapist has greater control over the composition of the modeling scene because the film or videotape can be reconstructed until the most desirable scene is produced. These media also permit the convenient use of multiple models, repeated observations of the same model, reuse of the films or videotapes with other persons, and self-administered treatment sessions. In addition to these advantages, there is the benefit of efficiency. More clients can be treated, and there is less demand on the time the professional spends with each client.

This article reviews research on the efficacy of videotape and film modeling, herein referred to as symbolic modeling (SM), in the treatment of clinical and clinical-like (i.e., analogue) problems. One purpose of this article is to review studies that test the efficacy of SM by comparing either SM alone, or SM combined with other treatment components, with control conditions. Studies that compare SM with other treatment techniques or components also are reviewed. Another purpose is to identify procedural and conceptual issues relevant to SM research. Finally, attention is given to methodological considerations that are critical to eventually discerning the effects of SM in a variety of therapeutic contexts and with various clinical groups.

Effects of Therapeutic Symbolic Modeling

Table 1 summarizes the studies that have investigated therapeutic SM. Note that the studies are grouped according to the behavior under investigation, beginning with phobias and followed by test anxiety, dental and medical stress, interpersonal skills, and other clinical problems. The second column in Table 1 specifies whether the subjects were adults or children and the target behavior under investigation. In the third column,

(*text continued on page 708*)

The authors are grateful to C. Steven Richards, Lawrence J. Siegel, and Joyce Winn for their valuable assistance with this article.

Requests for reprints should be sent to Mark H. Thelen, Department of Psychology, 210 McAlester Hall, University of Missouri, Columbia, Missouri 65211.

Table 1
Results of Symbolic-Modeling Studies

Study	Subjects and target behavior	Symbolic modeling versus control	Symbolic modeling versus comparison treatment	Generalization	Maintenance
		Phobias			
Bandura & Barab (1973)	Adults, snake fear	SM > irrelevant film on B, physiological, & SR*		Yes	Not assessed
Bandura, Blanchard, & Ritter (1969)	Adults, snake fear	SM + relaxation > no treatment on Att, B, & SR*	SM + relaxation > desensitization on SR & = desensitization on B; SM + relaxation < live modeling + participation on Att, B, & SR*	Yes	Yes (1 month)
Bandura & Menlove (1968)	Children, dog fear	SM > irrelevant film on B*		Yes	Yes (1 month)
Denney & Sullivan (1976)	Adults, spider fear	SM > no treatment on B & SR;* SM + relaxation > no treatment on B & SR*	SM > spider-only film on B & SR	Not assessed	Not assessed
Hill, Liebert, & Mott (1968)	Children, dog fear	SM > no treatment on B		Not assessed	Not assessed
Kornhaber & Schroeder (1975)	Children, snake fear	SM > no treatment on Att & B*		Not assessed	Not assessed
Lewis (1974)	Children, water fear	SM > irrelevant film on B; SM + guided participation > irrelevant film on B	SM < guided participation on B; SM + guided participation > guided participation on B; SM + guided participation > SM on B	Yes	Not assessed
Lira, Nay, McCullough, & Etkin (1975)	Adults, snake fear	SM > no treatment on Att & = no treatment on B & SR	SM > snake-only film on Att & = snake-only film on B & SR; SM < rehearsal on B & = rehearsal on Att & SR	Not assessed	Yes (2 months)
Morris, Spiegler, & Liebert (1974)	Adolescents, snake fear	SM > irrelevant film on SR*	SM > snake-only film on SR*	Not assessed	Not assessed

Table 1 (*continued*)

Study	Subjects and target behavior	Symbolic modeling versus control	Symbolic modeling versus comparison treatment	Generalization	Maintenance
		Phobias (*continued*)			
Spiegler, Liebert, McMains, & Fernandez (1969)					
Experiment 1	Adults, snake fear	SM > no treatment on B*[a]		Not assessed	Yes (1 month)
Experiment 2	Adults, snake fear	SM > no treatment on B		Not assessed	Not assessed
Experiment 3	Adults, snake fear		SM + relaxation > relaxation on B	Yes	Yes (1 week)
Weissbrod & Bryan (1973)	Male children, snake fear		SM > SM with snake replica on B	Not assessed	Yes (2 weeks)
		Test anxiety			
Andrews (1973)	Adults, test anxiety	SM + relaxation > no treatment on B & = no treatment on SR		Not assessed	Not assessed
Jaffe & Carlson (1972)	Adults, test anxiety	SM > no treatment on B & SR*		Not assessed	Not assessed
Malec, Pack, & Watkins (1976)	Adults, test anxiety	SM > no treatment on SR*		Not assessed	Not assessed
Mann (1972)	Adults, test anxiety	SM > no treatment on B & SR; SM + relaxation > no treatment on B & SR	SM = SM + relaxation on B & SR	Not assessed	Yes (6 weeks)
		Dental–medical stress			
Machen & Johnson (1974)	Preschoolers, dental fear	SM > control on B	SM = desensitization on B	Not assessed	Not assessed
Melamed, Hawes, Heiby, & Glick (1975)	Children, dental fear	SM > irrelevant film on B & staff rating & = irrelevant film on physiological & SR		Not assessed	Not assessed
Melamed & Siegel (1975)	Children, hospital fear	SM > irrelevant film on physiological, B, SR,* & parent ratings*		Not assessed	Yes* (3 weeks)
Melamed, Weinstein, Hawes, & Katin-Borland (1975)	Children, dental fear	SM > irrelevant task on B & staff rating & = irrelevant task on SR		Not assessed	Not assessed

(*table continued*)

Table 1 (*continued*)

Study	Subjects and target behavior	Symbolic modeling versus control	Symbolic modeling versus comparison treatment	Generalization	Maintenance
		Dental–medical stress (*continued*)			
Shaw & Thoresen (1974)	Adults, dental fear	SM + relaxation > assessment only on B, SR,* & Att*	SM + relaxation > audiotape discussion on B, Att,* & SR;* SM + relaxation > desensitization on Att* & = desensitization on B & SR	Not assessed	Yes (3 months)
Vernon (1973)	Children, anesthesia fear	SM > control on B* & = control on parent rating		Not assessed	Yes* (4 weeks)
Vernon (1974)	Children, inoculation fear	SM > control on B* & = control on staff rating		Not assessed	Not assessed
Vernon & Bailey (1974)	Children, anesthesia fear	SM > control on B & = control on staff rating		Not assessed	Not assessed
Wroblewski, Jacob, & Rehm (1977)	Adults, dental fear	SM = placebo on B & SR	SM < SM + relaxation on B* & = SM + relaxation on SR	Not assessed	Yes* (2 weeks)
		Interpersonal skills[b]			
Burrs & Kapche (Note 1)	Adult psychiatric patients, social skills deficit	SM + 3 components = no treatment on SR & staff rating		No	No (1 month[e])
Curran (1975)	College students, dating anxiety	SM + 4 components > waiting list on B & = waiting list on SR	SM + 4 components = desensitization on B & SR; SM + 4 components > relaxation on B & = relaxation on SR	Not assessed	Not assessed
Curran & Gilbert (1975)	College students, dating anxiety	SM + 6 components > minimal contact on B* & SR	SM + 6 components > desensitization on B* & = desensitization on SR	Not assessed	Yes (6 months)
Curran, Gilbert, & Little (1976)	College students, dating anxiety		SM + 5 components > sensitivity training on B & SR*	Not assessed	Not assessed
Eisler, Hersen, & Miller (1973)	Adult male psychiatric patients, unassertiveness	SM + rehearsal > test–retest on B*	SM + rehearsal > rehearsal on B*	Not assessed	Not assessed

Table 1 (*continued*)

Study	Subjects and target behavior	Symbolic modeling versus control	Symbolic modeling versus comparison treatment	Generalization	Maintenance
		Interpersonal skills[b] (*continued*)			
Evers & Schwartz (1973)	Preschoolers, social isolation		SM < SM + praise on B*	Not assessed	No (4 weeks)
Evers-Pasquale & Sherman (1975)	Preschoolers, social isolation	SM > control film on B		Not assessed	Yes (4 weeks)
Galassi, Galassi, & Litz (1974)	College students, unassertiveness	SM + 3 components > assessment only on B* & SR		Not assessed	Not assessed
Galassi, Kostka, & Galassi (1975)	Follow-up of Galassi et al. (1974)			Not assessed	Yes* (1 year)
Goldstein et al. (1973) Experiment 1	Adult neurotics, dependence	SM > assessment only on B*		Not assessed	Not assessed
Experiment 3	Adult schizophrenics, independence	SM > assessment only on B; SM + instructions > assessment only on B	SM = instructions on B; SM = SM + instructions on B; SM + instructions = instructions on B	Not assessed	Not assessed
Gottman (1977)	Preschoolers, social isolation	SM = control film on B & status		Not assessed	No[d] (6 weeks)
Gottman, Gonso, & Schuler (1976)	Children, low sociometric status	SM + 2 components > attention control on status* & = attention control on B		Not assessed	Yes*[d] (6 weeks)
Gutride, Goldstein, & Hunter (1973)	Adult psychiatric patients, social interaction	SM + 3 components > assessment only on B* & SR*		No	Not assessed
Hersen, Eisler, & Miller (1974)	Adult psychiatric patients, unassertiveness	SM + 2 components > test–retest on B*	SM + 2 components > rehearsal on B*	Yes*	Not assessed

(*table continued*)

Table 1 (*continued*)

Study	Subjects and target behavior	Symbolic modeling versus control	Symbolic modeling versus comparison treatment	Generalization	Maintenance
		Interpersonal skills[b] (*continued*)			
Hersen, Eisler, Miller, Johnson, & Pinkston (1973)	Adult psychiatric patients, unassertiveness	SM + 2 components > test–retest on B* & = test–retest on SR; SM + rehearsal > test–retest on B* & = test–retest on SR	See Footnote e; SM + 2 components > SM + rehearsal on B* & = SM + rehearsal on SR	Not assessed	Not assessed
Jaffe & Carlson (1976)	Adult male psychiatric patients, asocial behavior	SM + rehearsal > attention control on B* & staff ratings*	SM + rehearsal = instructions + rehearsal on B & < instructions + rehearsal on staff ratings*	Yes*	Yes* (1 month)
Jakibchuk & Smeriglio (1976)	Preschoolers, social isolation	SM + self-guiding comments > control film on B; SM + self-guiding comments > assessment only on B; SM + narration = control film on B; SM + narration = assessment only on B[f]	SM + self-guiding comments > SM + narration on B[f]	Not assessed	Yes (3 weeks)
Keller & Carlson (1974)	Preschoolers, social isolation	SM > control film on B		Not assessed	No (3 weeks)
McFall & Twentyman (1973)	College students, unassertiveness	SM + rehearsal > assessment only on SR	SM + rehearsal = audiotaped modeling + rehearsal on SR	Yes	Yes (2 weeks)
O'Connor (1969)	Preschoolers, social isolation	SM > control film on B		Not assessed	Not assessed
O'Connor (1972)	Preschoolers, social isolation	SM > control film on B; SM + shaping > control film on B	SM > shaping on B; SM + reinforcement > reinforcement on B; SM + reinforcement = SM on B	Not assessed	Yes (3 weeks)
Rathus (1973)	College students, unassertiveness	SM + 2 components > placebo on B & SR;* SM + 2 components > assessment only on B & SR*		Not assessed	Not assessed

Table 1 (*continued*)

Study	Subjects and target behavior	Symbolic modeling versus control	Symbolic modeling versus comparison treatment	Generalization	Maintenance
		Interpersonal skills[b] (*continued*)			
Thelen, Fry, Dollinger, & Paul (1976)	Male delinquents, interpersonal behavior	SM + rehearsal > control tapes on staff ratings & = control tapes on teacher ratings		Not assessed	No (2 weeks)
		Other clinical problems			
Baker, Udin, & Vogler (1975)	Adult males, alcoholism		SM < self-confrontation + behavioral counseling on B	Not assessed	No (6 weeks[g])
Reeder & Kunce (1976)	Adults, heroin addiction		SM + discussion > lecture + discussion on B	Not assessed	Yes[h]
Thomas (1974)	Children, classroom attention	SM > no treatment on B		Not assessed	Not assessed
VanCamp (1972)	Adults, marital conflict	SM = practice on B & SR		Not assessed	Not assessed
Wincze & Caird (1976)	Adult females, sex anxiety	SM > waiting list on SR*	SM < systematic desensitization on SR*	Not assessed	Not assessed

Note. SM = symbolic modeling, B = behavioral measure, SR = self-report measure excluding attitude measures, and Att = attitude measure; > indicates greater improvement, = indicates no differences, and < indicates less improvement.

[a] SM without narration group = no treatment on B; SM + narration group > no treatment on B.

[b] When a condition consisted of two or more components, in addition to SM, only the number of components is indicated. The following components were used in one or more of the studies: instructions, rehearsal, feedback, homework, relaxation, discussion, reinforcement, and communication training.

[c] Similar results were obtained at a 3-month follow-up.

[d] Posttest data were collected 6 weeks after intervention.

[e] SM + rehearsal + instructions > rehearsal + instructions on B* & = rehearsal + instructions on SR; SM + rehearsal + instructions > rehearsal on B* & = rehearsal on SR; SM + rehearsal > rehearsal on B* & = rehearsal on SR; SM + rehearsal < rehearsal + instructions on some B measures and > rehearsal + instructions on other B measures, & = rehearsal + instructions on SR.

[f] Because of its importance in this study, narration is reported as a separate component. It is not considered as a separate component for the other studies.

[g] Similar results were obtained at a 6-month follow-up.

[h] Posttest data were collected 90 days and 180 days after intervention.

* Results were mixed.

results are reported that involved a no-treatment, attention, or placebo control. The treatment was either SM alone or SM plus one or more treatment components (e.g., instructions or behavioral rehearsal). Column 4 lists results involving a comparison treatment that was composed of one or more components. As with column 3, the SM condition involved either SM alone or SM plus one or more additional components. The last two columns of Table 1 contain information pertinent to generalization and maintenance—whether or not these factors were assessed and, if so, whether the results were affirmative or negative. It was also deemed useful to specify the time interval between the treatment intervention and the assessment of maintenance.

The results, as they are summarized in Table 1, pertain to between-groups differences, usually based on posttest differences or differences in change scores between groups. Also, it should be noted that in all studies the symbol *greater than* (>) is meant to indicate greater improvement in the clinical behavior. For example, the first study listed is by Bandura and Barab (1973). The table shows that SM was greater than an irrelevant film. This should be interpreted to mean that the SM group improved more than the control group, that is, the SM group showed greater fear reduction. An equal sign (=) indicates no differences between groups.

It should be noted that in this portion of the article no attempt is made to qualify the results because of methodological considerations. Methodology issues are discussed in a later section.

Phobias

As can be seen in Table 1, research on phobias has been done with both adults and children; snake fear is the most common problem studied. The third column in the table reveals that SM, both with or without added components, had a positive treatment effect in comparison with the no-treatment and irrelevant-film control conditions. The fourth column reveals that SM compares favorably with other treatment components, especially if SM is combined with relaxation or guided participation. It also is apparent from the last two columns of Table 1 that, when assessed, generalization and maintenance effects were obtained.

Text Anxiety

Four studies have been reported that appraised the effects of SM, often involving vicarious systematic densensitization, on test-anxious adults. Table 1 reveals that generally positive effects have been obtained for SM when compared with control conditions; however, no research has compared SM with other treatment components. None of these studies reported an assessment of generalization, and only one study (Mann, 1972) assessed the maintenance of the treatment effects. Mann found that changes were maintained 6 weeks after treatment.

Dental–Medical Stress

Symbolic modeling has been used to alleviate stress and disruptiveness in dental and medical patients, typically among people who had no prior dental work or surgery. Prior to the actual dental work or surgery, the modeling film or videotape is shown to each treatment subject. Looking at the column of Table 1 that compares SM with control groups, it can be seen that all but one of the studies (Shaw & Thoresen, 1974) used SM alone, and the studies typically did not involve other components. It appears that SM often is effective in comparison with controls according to behaviorally based measures, but the results are less clear for measures obtained from staff ratings, peer ratings, and self-report.

It is apparent from column 4 that only three studies (Machen & Johnson, 1974; Shaw & Thoresen, 1974; Wroblewski, Jacob, & Rehm, 1977) have investigated the effect of SM relative to other treatment components. Given the mixed findings of these few studies, a definitive statement is not possible.

Generalization as such has not been assessed by researchers in this area; however, the dependent measure was often taken in a natural setting, so that additional measures to assess generalization may not have been necessary. Four studies (Melamed & Siegel,

1975; Shaw & Thoresen, 1974; Vernon, 1973; Wroblewski et al., 1977) appraised maintenance of treatment effects for periods ranging from 3 weeks to 3 months, and all of these studies found at least some indication of maintenance.

Interpersonal Skills

Symbolic modeling has been used to facilitate assertiveness, social interaction, and the development of other appropriate social skills. This area of research contains the largest number of studies involving SM. Studies have been done with children, adolescents, college students, and adult psychiatric patients. A summary of the results depicted in the third column of Table 1 shows that SM alone was generally more effective than the control conditions, although two studies found no differences between SM and control conditions on social interaction in socially isolated preschoolers (Gottman, 1977; Jakibchuk & Smeriglio, 1976). With five exceptions SM plus other treatment components, often three or four in number (see Footnote b in Table 1), typically was greater in effect than the control condition. In three of these five studies, the failure to obtain significant differences between a treatment condition including SM and a control condition was based on self-report measures (Curran, 1975; Hersen, Eisler, Miller, Johnson, & Pinkston, 1973; Burrs & Kapche, Note 1).

Of the studies that involved SM plus components compared with a control group, nearly all (13 of 15 studies) contained rehearsal. Five studies contained an instruction component, 5 a feedback component, 4 a discussion component, 3 a homework component, 2 a reinforcement component, and 1 a communication training component. Of the 13 studies that contained SM plus rehearsal, all but 5 also contained other components. Of these 5 studies, 3 found SM plus rehearsal to be more effective than a control condition according to behavioral measures (Eisler, Hersen, & Miller, 1973; Hersen et al., 1973; Jaffe & Carlson, 1976). McFall and Twentyman (1973) reported that SM plus rehearsal resulted in greater improvement than did the control condition, based on a self-report measure, but Hersen et al. found no differences in improvement with a self-report measure. Jaffe and Carlson and Thelen, Fry, Dollinger, and Paul (1976) found SM plus rehearsal to be more effective than control conditions according to staff ratings, but the latter study reported no differences on teacher ratings. Since a number of studies found SM (without additional components) to have significant effects, the results in the above studies may have stemmed from SM alone. No research has independently manipulated SM and rehearsal to appraise their separate and joint effects. Definitive statements about the other components are even more difficult to make because there are few studies and these nearly always involved a multicomponent treatment package.

With the exception of the study by Goldstein et al. (1973), the studies that compared SM with other treatments involved SM plus one or more additional components. The two components most frequently combined with SM were rehearsal (eight studies) and instructions (seven studies). Of the 17 group comparisons involving SM plus one or more additional components in which a behavioral measure was used, the condition that included SM was greater in effect than the comparison treatment in 11 instances and equal to the comparison treatment in 6 instances. However, the condition that contained SM was greater in effect than the comparison treatment in only 1 (Curran, Gilbert, & Little, 1976) of the 10 instances that involved a self-report measure. It appears that SM plus components compares favorably with other treatments, based on behavioral measures, but differences on self-report measures seldom have been found. One explanation for this consistent finding (differences on behavioral but not on self-report measures) is that the behavioral measures typically are obtained during role-playing situations that are similar if not the same as those used in the treatment session. On the other hand, the self-report measures are generally more global measures of social skills. Hersen & Bellack (1976) proposed that there is an "attitudinal lag" between the rapid behavior changes in social skills and subsequent self-assessments.

Also shown in the fourth column of Table 1 are five studies in which the comparison treatment included SM. These studies indicate that SM plus praise is more effective than SM alone in increasing social interaction among socially isolated children (Evers & Schwartz, 1973). However, O'Connor (1972) found that shaping did not make a similar additive contribution to SM alone. Instructions, when included in a treatment package including SM plus rehearsal, facilitated the development of assertive behaviors more than did just SM plus rehearsal (Hersen et al., 1973). However, SM plus instructions was equivalent to SM alone in increasing independent behavior (Goldstein et al., 1973). Also, Jakibchuk and Smeriglio (1976) reported that SM that included first-person, self-guiding comments was superior to SM that included narrative, third-person comments in the sound track. This study is discussed further in a later section. Because of insufficient research on any one component and differences in the dependent measures and populations employed, definitive statements are not possible.

In this area of research generalization usually has not been assessed. Of those studies that assessed generalization, 3 found that SM, at least as one component, facilitated generalization (Hersen, Eisler, & Miller, 1974; Jaffe & Carlson, 1976; McFall & Twentyman, 1973), whereas 2 studies failed to find such effects (Gutride, Goldstein, & Hunter, 1973; Burrs & Kapche, Note 1). However, as with the dental–medical research, studies involving social isolation in preschoolers have assessed the subjects' behavior in natural settings, that is, during children's interactions at nursery school. Generalization to home or other settings has not been assessed. The results concerning maintenance are also very mixed. Of the 13 studies that assessed maintenance, 8 found some evidence for maintenance, and 5 obtained negative results.

Other Clinical Problems

The section of Table 1 labeled Other Clinical Problems contains five studies that focused on problems not encompassed by the previous sections. Since no problem area was investigated more than once, it is impossible to make any definitive observation with respect to the influence of SM. These studies are mainly included for the sake of comprehensiveness and to suggest possible directions for future research.

Conceptual and Process Considerations

Attention is given in this section to variables and processes that have been studied within the context of SM research and that have relevance to theoretical conceptualizations of modeling. Attending to process considerations may facilitate theorizing regarding modeling and further refinement of the applications of SM.

Model–Observer Similarity

The importance of model–observer similarity has drawn the attention of many researchers. Two variables in particular, model age in relation to observer age and the presentation of a coping versus a mastery model, have received considerable attention.

Bandura and Barab (1973) found that adults showed as much snake-fear reduction after observing a child model as subjects who observed a peer model. Fear measures included behavioral, self-report, and physiological measures. Similarly, Weissbrod and Bryan (1973) exposed children to a same-age or a younger model. They found no difference between the two groups on a behavioral approach to snakes measure. Neither of the above two studies contained a model who was older than the subjects. Kornhaber and Schroeder (1975) addressed this question by exposing snake-fearing children to either a child or an adult model. Those subjects who observed a child model demonstrated greater fear reduction on a behavioral avoidance test than those who observed an adult model, but there were no differences on an attitude measure. Furthermore, Kornhaber and Schroeder found that children who observed an adult model did not show greater fear reduction than did a no-treatment control group. Although these studies are limited in number, they are reasonably consistent. With adult subjects, it appears that observing a same-age or child model facilitates fear reduction,

Similarly, children experience the same degree of fear reduction when exposed to peer-age or younger models. However, the observation of an adult model has minimal effects on children. Children may see older models as superior and more capable and therefore may not imitate their approach behavior in a fearful situation. On the other hand, when adults observe children or when children observe younger children, they may be motivated to perform the approach response (Bandura & Barab, 1973). Bandura and Barab reported that adult subjects who observed an adult model showed a positive correlation between fear extinction (based on autonomic responses) and behavioral improvement, which they interpreted as vicarious extinction. In contrast, the adult subjects who observed child models demonstrated little relationship between fear extinction and behavioral improvement, which suggests that motivational factor served to influence approach behavior. In other words, "If that kid can do it so can I." No related research has been done on the effects of model–subject age differences when both model and subject are adults.

The question of the effects of a coping model versus a mastery model is a very complicated one that touches on many of the studies on SM. Many of the studies reviewed used a coping model, but did not manipulate the coping–mastery variable. Additionally, there were variations of the coping–mastery variable. For example, Curran (1975) presented an inappropriate model followed by a competent model.

In a study with snake-avoiding college females, Meichenbaum (1971) suggested that the coping–mastery variable might relate to the broader question of model similarity, which has been shown to enhance imitation (Rosekrans, 1967). In the Meichenbaum study, the coping models displayed initially hesitant behavior, but later fearlessly interacted with the target snake. The mastery models were fearless throughout the modeling sequences, with each model confidently handling the snake. Meichenbaum found that the subjects who observed the coping models displayed significantly greater improvement than subjects in the mastery condition. A similar procedure was used to alleviate interview anxiety in psychiatric patients (Bruch, 1975). Bruch's results were not as consistent as Meichenbaum's, but he did report improvement of subjective anxiety according to self-ratings among the coping subjects as compared with the mastery subjects. Working in the area of anxiety associated with surgery, Vernon (1974) reported that subjects who observed a model who showed some degree of pain were rated as experiencing less pain subsequently than were subjects who observed a model who did not demonstrate pain. Perhaps a study that fails to demonstrate effects is as important as affirmative studies. Lira, Nay, McCullough, and Etkin (1975) reported that adult female subjects who observed a fearless "expert" model were uninfluenced by that model as compared with a control group, based on behavioral and self-report measures. Meichenbaum (1971) acknowledged, however, that the greater effectiveness of the coping model in his study could have stemmed from the modeling of coping techniques that were an inherent feature of the coping conditions, rather than from the perceived similarity between observer and model per se. In consideration of this point, Kornhaber and Schroeder (1975) presented models who did not demonstrate deep breathing, which was central to Meichenbaum's coping model. With this confound removed, Kornhaber and Schroeder found greater change in evaluative attitudes toward snakes among those who observed a fearful model as compared with those who observed a fearless model, but they obtained no differences on a behavioral measure. Also, Jaffe and Carlson (1972) found no differences on behavioral and self-report measures of test anxiety between adults who observed a calm model and adults who observed an anxious model. These findings are generally consistent with Kornhaber and Schroeder's (1975) suggestion that with model coping skills removed, the greatest effect of a coping model may be on attitudes rather than on overt behavior.

A number of other researchers have adopted the Meichenbaum procedures in their applications of SM to clinical problems. The more focal problems of phobias and stress associated with dental and medical procedures

have been responsive to the coping manipulations. Lewis (1974) found water-phobic subjects to be significantly improved at the posttreatment assessment after they observed a coping model. Hill, Liebert, and Mott (1968) and Spiegler, Liebert, McMains, and Fernandez (1969) combined the mastery and coping procedures by depicting an "expert" model who demonstrated competent dog or snake handling to an initially apprehensive (i.e., coping) model. The second model then imitated the first model. Both research groups found moderate support for the combined modeling approaches, based on behavioral measures. In her studies of dental anxiety (Melamed, Hawes, Heiby, & Glick, 1975; Melamed, Weinstein, Hawes, & Katin-Borland, 1975) and anxiety associated with surgery (Melamed & Siegel, 1975), Melamed used coping models to successfully reduce the disruptive behavior of children undergoing these procedures. Except for Gottman (Gottman, 1977; Gottman, Gonso, & Schuler, 1976), who obtained weak effects, most studies have reported positive effects in using a coping model with interpersonal problems (Evers & Schwartz, 1973; Evers-Pasquale & Sherman, 1975; Jakibchuk & Smeriglio, 1976; O'Connor, 1969, 1972; Thelen et al., 1976).

Narration

Bandura (1977) has emphasized the importance of attentional and imaginal-verbal cognitive processes in modeling. In general, four subprocesses of modeling are emphasized by Bandura. In order to imitate, the observer must first attend to the critical model behavior. Second, the cognitive processes of imagery labeling and verbal labeling that are sustained over time by rehearsal facilitate later imitation of a model when external modeling cues are not present. Third, in order to overtly imitate a model, a person must have the motor reproduction abilities that are particularly relevant to more complex motor behavior. Last, although modeled behavior can be acquired by observation alone, reinforcement is critical if observed behavior is to be performed and this performance is to be continued over time.

Some aspects of the SM research reviewed in this article are relevant to these conceptualizations. One variable that has received considerable attention is narration, or a description of the model behavior concurrent with model performance. Based on current theorizing, narration should facilitate attention to the model and verbal labeling of the critical model behavior and thereby should increase the effectiveness of modeling interventions. Jakibchuk and Smeriglio (1976) reported that SM plus narration failed to significantly affect the social behavior of isolated preschool children in comparison with two control groups. However, researchers have reported that SM plus narration diminishes snake fears in adults (Spiegler et al., 1969), social isolation among preschoolers (Evers-Pasquale & Sherman, 1975; Keller & Carlson, 1974; O'Connor, 1969, 1972), and anesthesia fear among children (Vernon & Bailey, 1974). Since SM alone often has a favorable effect, it is of questionable value to demonstrate that SM plus narration has a favorable effect. Perhaps the greater question is whether narration has an incremental effect on SM alone. On this question Morris, Spiegler, & Liebert (1974) found that SM plus narration was equal to SM alone, based on a self-report measure of snake fear among adolescents. No other studies directly compared SM plus narration with SM alone.

Although the effect of adding narration to SM is uncertain, Jakibchuk and Smeriglio (1976) reported a timely study in which they compared narration with the influence of self-guiding model comments on the social isolation behavior of preschool children. They found that SM plus self-guiding comments by the model had a significant effect in comparison with two control conditions and in comparison with an SM plus narration condition. These findings are consistent with the earlier work of Meichenbaum (1971), who reported that a model's self-verbalization facilitated reduction of snake fear among female undergraduates. It may be that self-verbalization facilitates the conversion of the model's behavior into perceptual-cognitive images, as described by Bandura (1977). Perhaps model self-verbalizations are especially facilitative when the model behavior is relatively complex; however, there is little systematic research relevant to this point.

Context and Complexity of Model Behavior

The nature of the model context may be critical to attentional processes. It is suggested that the model context needs to be sufficiently simple to insure attention to the critical model behaviors, but it also should contain enough contextual cues to facilitate generalization to other settings. Furthermore, as the complexity of the target behavior increases, there may be greater need to both simplify and amplify the target behaviors. An example of such simplification of complex target behaviors is Hersen et al.'s (1973) reliance on eight response modalities characteristic of assertive responses. These included a lengthy reply, a request for behavior change, consistent eye contact, a refusal to comply with unreasonable requests, a fully audible speaking voice, an assertive affect, a quick latency of response, and overall assertiveness. Elaborative or focusing narration, self-guiding comments, or instructions may be used in conjunction with simplified target behavior to further increase the subject's attentiveness to relevant materials. By these means, the subject's attention is directed to the subtleties of the interpersonal interactions, which may otherwise go unnoticed in daily naturalistic modeling.

In the course of developing modeling films for interpersonal skills, it is important to avoid overwhelming the observer. Even with simplification and elaboration it is possible to overload the observer with too many significant points of focus. Burrs and Kapche (Note 1) failed to achieve successful results using SM with psychotic inpatients. In his criticism of their study, Goldstein (1973) cited their research as an attempt to cover too much ground in short videotapes. However, when combined with role playing, SM may be effective with more complex skills, for example, requesting new behavior, whereas instructions may be adequate for less complex skills, for example, eye contact (Eisler et al., 1973; Hersen et al., 1974; Hersen et al., 1973).

Uncertainty, Arousal, and Model Warmth

Other considerations are also very likely to be relevant to the matter of attention to and imitation of the model. For example, Yussen & Levy (1975) found that subjects who observed a warm live model attended to the model and recalled the model's behavior more than did subjects who observed a neutral live model. Others have suggested that uncertainty in the observer may increase attention to and subsequent imitation of a model (Marlatt, 1972; Thelen, Paul, Dollinger, & Roberts, 1978). If this is the case, it is likely that individuals will profit most from SM at those times when they actively seek information regarding appropriate behavior, for example, shortly before a stressful medical procedure. On the other hand, if an observer experiences extreme anxiety in anticipation of a stressful event or if the model's behavior creates a great deal of anxiety in the observer, the observer may avoid attending to the model.

Retention

As reflected in the methodologies of much of the SM research, especially in the failure to routinely collect data on maintenance, there is relatively little material relevant to the question of retention. The behavioral rehearsal and practice components that were so frequently used in the interpersonal skills research might be expected to facilitate not only immediate imitation of the model but also the retention of the modeled behavior (Bandura, 1977). However, there is no research that documents the influence of rehearsal and practice following observation of a symbolic model on the retention of behavior. Research by Thelen, Fryrear, and Rennie (1971) with a nonclinical target behavior (i.e., standards of self-reward) suggests that the influence of observing a model can have long-term effects.

Model Consequences

If shown to be useful, model consequences can easily be incorporated as part of SM. Also, model consequences have been extensively researched (Thelen & Rennie, 1972) and are important to the theorizing of Bandura (1977), who ascribed both informational and incentive functions to this variable. Many of the studies that focused on anxiety-related problems did not contain model consequences,

perhaps because of the assumption that the absence of negative consequences following the demonstration of the feared act in itself serves as positive model reinforcement. Model consequences were included in some of the studies on social isolation in preschoolers (Evers & Schwartz, 1973; Evers-Pasquale & Sherman, 1975; O'Connor, 1969, 1972), on assertion (McFall & Twentyman, 1973), and on dental and medical stress (Melamed, Hawes, Heiby, & Glick, 1975; Melamed, Weinstein, Hawes, & Katin-Borland, 1975). Although all of these studies showed some positive effects for SM, none of them manipulated the model-consequences variable. Therefore, this research does not establish that model consequences have an additive effect on the influence of SM alone.

Acquisition Versus Disinhibition

Yet another aspect of Bandura's (1977) theory of imitation is his distinction between two effects of modeling: disinhibition and the acquisition of behavior. A behavior that has previously been acquired may not be performed because of fear or anxiety. Observation of a model engaging in the behavior without negative consequences or with positive consequences may disinhibit the behavior or increase its frequency. New patterns or sequences of behavioral components also are exhibited by observers following observation of a model. In this manner novel behaviors are acquired.

It is potentially important to both application and theory to determine if SM primarily facilitates anxiety reduction and disinhibition or skills acquisition. Of course, it is possible that SM has an important influence in both areas. The treatment components that researchers have combined or compared with SM may give some clues as to their thinking regarding this matter. Research on the use of SM with phobias, test anxieties, and dental–medical stress used primarily relaxation and desensitization as additional components and as comparison treatments. One might infer that researchers see these problems as primarily involving anxiety and inhibition. In contrast, the components (e.g., rehearsal) used by most researchers of interpersonal problems reflect an assumption of a skills deficit. However, two recent studies raise the question as to the extent that skills deficits are the cause of some of the interpersonal problems studied. Schwartz and Gottman (1976) found that low-assertive college students were as able as moderate- and high-assertive college students to write out what they thought a good assertive response might be. Even more significantly, when asked to role play an assertive response for a hypothetical unassertive friend, the low-assertive students were as assertive as the moderate- and high-assertive students. Consistent with the above, Nietzel and Bernstein (1976) found that high-demand instructions made unassertive college students more assertive than did low-demand instructions. Both of the above studies suggest that assertive responses were in the behavioral repertoires of the low-assertive college students. However, since both studies were with college students, similar conclusions regarding other clinical populations are not possible. For example, it is likely that unassertive psychiatric inpatients have a more severe behavioral deficit than do unassertive college students (Hersen & Bellack, 1976).

Morris et al. (1974) have suggested that there are two components in phobia problems, one cognitive and the other emotional. The cognitive component involves information about alternative responses and expected outcomes that might ensue. The emotional aspect involves autonomic processes. Morris et al. suggested that the cognitive aspect should dominate in the anxiety of relatively normal subjects with snake fears because of their lack of experience and their faulty beliefs. They found that high school children with snake fears who observed a symbolic model showed more fear reduction than a control group, based on a cognitive (worry) measure but not on an emotionality (autonomic reactions) measure. Based on these findings, one might infer that SM is better suited to changing cognitive functioning and perhaps not so effective in changing autonomic processes. These findings are generally consistent with a study by Bandura and Menlove (1968), in which they showed that high-emotion-prone subjects in a multiple model condition showed less reduction in dog

fear than subjects who were not high-emotion-prone. Morris et al. worked with relatively normal subjects, who may be more similar to the moderate-emotion-prone subjects than to the high-emotion-prone subjects in the Bandura and Menlove study. One implication of these findings may be that SM is not as effective with more severe clinical problems that involve a strong anxiety component. Research is needed that addresses this question.

Multiple Models

Another facet of an SM treatment that is easily incorporated is the use of multiple models instead of the presentation of a single model. The purpose of using multiple models is to increase the odds that the observer will select at least one model to imitate, to provide multiple exposures to the target behavior, to increase the treatment stability, and to facilitate generalization of the behavior. Many of the studies reviewed in Table 1 employed multiple models, and this was especially likely in those studies that assessed generalization. However, only Bandura and Menlove (1968) actually manipulated the number of models. In their study the usual posttreatment, follow-up, and generalization comparisons failed to differentiate the dog-fearing children who observed multiple models from those who observed a single model. However, more children in the former group completed the behavioral approach test than did children in the single model group. Since the use of multiple models is readily incorporated in an SM treatment, and since multiple models logically seem to facilitate generalization, this question is clearly in need of research.

Length of Model Presentation

The length of the model presentation is another consideration that is potentially critical to the effectiveness of SM and on which the studies vary considerably. However, none of these studies systematically manipulated the length of the model presentation, and therefore little can be gleaned in this regard. A study by McGuire, Thelen, and Amolsch (1975) of self-disclosure showed that instructions were as effective as modeling when the audio model presentation was brief, but when the model presentation was longer, the audio model was more effective than were instructions. If SM is to become a treatment in general use, it is important to establish the optimal length of model presentation for various clinical problems and clinical populations.

Methodological Considerations

One could write an entire volume on methodological concerns in this area of research; only some of the more critical matters are discussed here. These points generally pertain to a number of studies rather than to one or two isolated studies. The following discussion concerns (a) subject selection, (b) defining the treatment or the independent variable, (c) control groups, (d) criterion measures of change, and (e) generalization and maintenance.

Subject Selection

No study can be better than the accuracy of the clinical population sampled. In the case of interpersonal skills problems, some researchers obtained subjects through psychology classes, which is probably not as desirable as obtaining subjects through newspaper advertisements (Little, Curran, & Gilbert, 1977). Psychology students are a relatively homogeneous population, and they are seldom completely naive concerning the experimental hypotheses. Inferences from studies that used college students to more diverse populations are highly risky. A strength in the research on dental–medical fears is that the subjects were drawn from the clinical setting in which the problem was demonstrated.

Another related problem is the tendency in SM research to use people who have relatively minor clinical problems. This is most apparent in the research on snake phobias. As Morris et al. (1974) suggested, the efficacy of SM with relatively normal snake phobics may not obtain when SM is applied to a more clinical population. There is a dire need for research on the effects of SM with more truly clinical and more disturbed populations.

Definition of the Independent Variable

The definition of the independent variable or the composition of the treatment is another

critical area. Although some of the studies compared SM alone with one or more control conditions, the prominent tendency, especially in the studies on interpersonal skills, has been to put SM with a basket of components. The net result is that one knows relatively little about the individual influence of SM as a separate component and relatively little about what SM might contribute to an intervention that contains one or more other components. Only a limited number of studies were designed to identify the influence of the specific components (cf. Goldstein et al., 1973; O'Connor, 1972).

Comparisons With Control Groups

If one wishes to attribute assessed changes to treatment variables, one must rule out the various effects that may be attributed to assessment, attention, placebo, or other demand characteristics. In general, the research on dental or medical stress and interpersonal skills controlled for the influence of assessment, attention, and placebo. But the research on phobias and test anxiety all too often contained only a no-treatment control, which does not allow one to rule out the influence of nontreatment variables. However, when the primary purpose of a study is to compare SM with other treatment techniques, an instruction/demand control may not be critical.

Many of the studies on interpersonal skills that used role playing (rehearsal) as part of the treatment package also used role playing to assess possible changes in interpersonal behavior, but did not control for practice in role playing. It is possible that any assessed changes in these studies stemmed from the increased competence of the treatment subjects at role playing, whether it pertained to assertiveness or to other behaviors. Therefore, if role playing is a part of the treatment and assessment, it is important to control for role playing as a general skill that can be learned.

Assessment

The matter of assessment in research on SM is both critical and complicated. A number of different methods have been used to assess the effect of the intervention, including self-report, physiological, and a variety of behavioral measures. Many researchers assessed behavioral changes in role-playing situations, whereas others assessed behavior by creating circumstances relevant to the behavior under investigation and unobtrusively observing the subjects' responses. Other behavioral measures that have been employed include direct observation in natural settings by trained observers and general ratings from informed acquaintances of the subjects.

Much of the research on SM contained a number of measures to assess the effects of the intervention, and most researchers in this area generally support the idea of multiple dependent measures. However, many researchers have pointed out the poor relationship between some of these dependent measures, as for example between behavioral and self-report measures of assertion (Rich & Schroeder, 1976) and of anxiety (Paul & Bernstein, 1973). Perusal of Table 1 reveals discrepant findings for relationships among the various dependent measures within a number of studies; that is, significant differences between groups were obtained on one type of measure (e.g., self-report) but not on another type (e.g., behavioral). Additionally, the large number of instances of mixed results on a given type of measure (as indicated by asterisks in Table 1) illustrates the frequent use of multiple behavioral and self-report measures. When a large number of measures are obtained, significant differences are more likely a function of chance.

A number of researchers have questioned the reliability and validity of the assessment methods used in SM research. For example, evidence has been presented indicating that the behavioral avoidance tests used to assess target anxiety in phobics are subject to a variety of uncontrolled situational and procedural influences (Bernstein & Nietzel, 1973). Bernstein and Nietzel argued that such influences as demand characteristics and instructional differences potentially obscure the relationship between treatment variables and anxiety. This is especially true when such influences result in the inclusion of less phobic subjects, who may respond more readily than

"true" phobics to posttreatment demands for approach behavior (Bernstein & Paul, 1971). Role-playing procedures used to assess assertiveness and heterosexual anxiety may be similarly influenced by instructionally mediated demand (Nietzel & Bernstein, 1976), thus limiting the validity of such procedures. The reliability and validity of most self-report measures of assertiveness have also been criticized (Rich & Schroeder, 1976), and such tests may be insensitive to treatment effects (see Table 1). Finally, Gottman (1977) criticized the naturalistic-observation assessment procedures employed in previous studies of SM with socially isolated preschoolers. These studies were criticized for the lack of detail in the coding systems employed, the failure to adequately control observer bias, the use of error-prone time sampling procedures, and the failure to control for interobserver reliability "decay" and observer "drift."

In summary, it seems to be important to incorporate multiple measures into an assessment of the effects of SM. However, there is a need for more serious consideration of assessment problems in future SM research. One last point deserves mention. Only three of the studies reviewed in Table 1 used a physiological measure to assess change. Even though there are problems with physiological measures, in that correlations between physiological measures are low and are sometimes inconvenient to obtain, greater use of such measures seems to be indicated. This would be especially true when treating anxiety-related problems and might eventually help to clarify the controversy concerning the relative roles of anxiety and skills deficit in the various areas studied.

Generalization and Maintenance

Since nearly all of this research, except that on dental and medical stress, was not carried out in a naturalistic setting, it is important to address the question of generalization. The second to last column in Table 1 demonstrates the relatively small percentage of studies that assessed generalization. And even some of the studies that assessed generalization did so in a way that leaves much to be desired. Recognizing that generalization is on a continuum, many of the researchers seem to have appraised generalization within a context that more closely approximated the treatment situation than the natural environment. For example, some of the phobia research appraised generalization by introducing a snake different from the one used for treatment. What does this tell us about the subject's response to a snake that they happen upon in their backyard on a Sunday afternoon? Similarly, a frequent method of assessing generalization in assertion research has been to introduce role-playing scenes that are different from those used for treatment. Surely this measure of generalization more closely approximates the treatment situation than the natural environment. The time has come to use generalization measures that appraise the subject's behavior outside the treatment setting and in settings more like the natural environment. The dental and medical research represents a clear exception to this problem. Since treatment and data collection typically occur within the natural environment, the question of generalization becomes less of a problem.

The question of maintenance has received about as little attention as generalization (see the last column of Table 1). When measures of the maintenance of effects were taken, they often were obtained within a few days after treatment. Even more critical is that these researchers typically did not assess the duration of the changes in the natural environment. In short, duration without generalization is of little value.

Conclusions

Based on the research reviewed in this article, it appears that the use of symbolic modeling as a treatment device, perhaps combined with other components, may have a promising future. If nothing else, the research to date suggests a great potential in this area. At this point, the need is for studies with more clinical or disturbed populations, for studies that systematically vary the treatment components, that carefully use multiple measures to assess change, and that appraise both generalization and maintenance in natural settings or in settings that closely approximate the natural environment.

Reference Note

1. Burrs, V., & Kapche, R. *Modeling of social behavior in chronic hospital patients*. Unpublished manuscript, California State College, Long Beach, 1969.

References

Andrews, J. A study of the effects of a filmed, vicarious, desensitization procedure in the treatment of manifest anxiety and test anxiety in community college students (Doctoral dissertation, Michigan State University, 1965). *Dissertation Abstracts International*, 1973, *33*, 5399A–5400A. (University Microfilms No. 66-351)

Baker, T. B., Udin, H., & Vogler, R. The effects of videotaped modeling and self-confrontation on the drinking behavior of alcoholics. *International Journal of the Addictions*, 1975, *10*, 779–793.

Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.

Bandura, A. *Social learning theory*. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Bandura, A., & Barab, P. Processes governing disinhibitory effects through symbolic modeling. *Journal of Abnormal Psychology*, 1973, *82*, 1–9.

Bandura, A., Blanchard, E., & Ritter, B. Relative efficacy of desensitization and modeling approaches for inducing behavioral, affective, and attitudinal changes. *Journal of Personality and Social Psychology*, 1969, *13*, 173–199.

Bandura, A., & Menlove, F. Factors determining vicarious extinction of avoidance behavior through symbolic modeling. *Journal of Personality and Social Psychology*, 1968, *8*, 99–108.

Bernstein, D. A., & Nietzel, M. T. Procedural variation in behavioral avoidance tests. *Journal of Consulting and Clinical Psychology*, 1973, *41*, 165–174.

Bernstein, D. A., & Paul, G. L. Some comments on therapy analogue research with small animal "phobias." *Journal of Behavior Therapy and Experimental Psychiatry*, 1971, *2*, 225–237.

Bruch, M. Influence of model characteristics on psychiatric inpatients' interview anxiety. *Journal of Abnormal Psychology*, 1975, *84*, 290–294.

Curran, J. P. Social skills training and systematic desensitization in reducing dating anxiety. *Behaviour Research and Therapy*, 1975, *13*, 65–68.

Curran, J. P., & Gilbert, F. S. A test of the relative effectiveness of a systematic desensitization program and an interpersonal skills training program with date anxious subjects. *Behavior Therapy*, 1975, *6*, 510–521.

Curran, J. P., Gilbert, F. S., & Little, L. M. A comparison between behavioral replication training and sensitivity training approaches to heterosexual dating anxiety. *Journal of Counseling Psychology*, 1976, *23*, 190–196.

Denney, D. R., & Sullivan, B. J. Desensitization and modeling treatments of spider fear using two types of scenes. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 573–579.

Eisler, R. M., Hersen, M., & Miller, P. M. Effects of modeling components of assertive behavior. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, *4*, 1–6.

Evers, W., & Schwartz, J. Modifying social withdrawal in preschoolers: The effects of filmed modeling and teacher praise. *Journal of Abnormal Child Psychology*, 1973, *1*, 248–256.

Evers-Pasquale, W., & Sherman, M. The reward value of peers: A variable influencing the efficacy of filmed modeling in modifying social isolation in preschoolers. *Journal of Abnormal Child Psychology*, 1975, *3*, 179–189.

Galassi, J. P., Galassi, M. D., & Litz, M. C. Assertive training in groups using video feedback. *Journal of Counseling Psychology*, 1974, *21*, 390–394.

Galassi, J. P., Kostka, M. D., & Galassi, M. D. Assertive training: A one year follow-up. *Journal of Counseling Psychology*, 1975, *22*, 451–452.

Geer, J., & Turtletaub, A. Fear reduction following observation of a model. *Journal of Personality and Social Psychology*, 1967, *6*, 327–331.

Goldstein, A. P. *Structured learning therapy: Toward a psychotherapy for the poor*. New York: Academic Press, 1973.

Goldstein, A. P., et al. The use of modeling to increase independent behavior. *Behaviour Research and Therapy*, 1973, *11*, 31–42.

Gottman, J. The effects of a modeling film on social isolation in preschool children: A methodological investigation. *Journal of Abnormal Child Psychology*, 1977, *5*, 69–78.

Gottman, J., Gonso, J., & Schuler, P. Teaching social skills to isolated children. *Journal of Abnormal Child Psychology*, 1976, *4*, 179–197.

Gutride, M., Goldstein, A., & Hunter, G. The use of modeling and role playing to increase social interaction among asocial psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1973, *40*, 408–415.

Hersen, M., & Bellack, A. S. Social skills training for chronic psychiatric patients: Rationale, research findings, and future directions. *Comprehensive Psychiatry*, 1976, *17*, 559–580.

Hersen, M., Eisler, R., & Miller, P. An experimental analysis of generalization in assertive training. *Behaviour Research and Therapy*, 1974, *12*, 295–310.

Hersen, M., Eisler, R., Miller, P., Johnson, M., & Pinkston, S. Effects of practice, instructions, and modeling on components of assertive behavior. *Behaviour Research and Therapy*, 1973, *11*, 443–451.

Hill, J. H., Liebert, R., & Mott, D. Vicarious extinction of avoidance behavior through films: An initial test. *Psychological Reports*, 1968, *22*, 192.

Jaffe, P. G., & Carlson, P. M. Modelling therapy for test anxiety: The role of model affect and consequences. *Behaviour Research and Therapy*, 1972, *10*, 329–339.

Jaffe, P. G., & Carlson, P. M. Relative efficacy of modeling and instructions in eliciting social behavior from chronic psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 200–207.

Jakibchuk, Z., & Smeriglio, V. L. The influence of symbolic modeling on the social behavior of pre-

school children with low levels of social responsiveness. *Child Development*, 1976, *47*, 838–841.

Keller, M., & Carlson, P. The use of symbolic modeling to promote social skills in preschool children with low levels of social responsiveness. *Child Development*, 1974, *45*, 912–919.

Kornhaber, R., & Schroeder, H. Importance of model similarity in extinction of avoidance behavior in children. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 601–607.

Lewis, S. A comparison of behavior therapy techniques in the reduction of fearful avoidance behavior. *Behavior Therapy*, 1974, *5*, 648–655.

Lira, F., Nay, W., McCullough, J., & Etkin, M. Relative effects of modeling and role playing in the treatment of avoidance behaviors. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 608–618.

Little, L. M., Curran, J. P., & Gilbert, F. S. The importance of subject recruitment procedures in therapy analogue studies on heterosexual-social anxiety. *Behavior Therapy*, 1977, *8*, 24–29.

Machen, J., & Johnson, R. Desensitization, model learning, and the dental behavior of children. *Journal of Dental Research*, 1974, *53*, 83–87.

Malec, J., Pack, T., & Watkins, J. T. Modeling with role playing as a treatment for test anxiety. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 679.

Mann, J. Vicarious desensitization of test anxiety through observation of videotaped treatment. *Journal of Counseling Psychology*, 1972, *19*, 1–7.

Marlatt, G. A. Task structure and the experimental modification of verbal behavior. *Psychological Bulletin*, 1972, *78*, 335–350.

McFall, R. M., & Twentyman, C. T. Four experiments on the relative contributions of rehearsal, modeling, and coaching to assertion training. *Journal of Abnormal Psychology*, 1973, *81*, 199–218.

McGuire, D., Thelen, M. H., & Amolsch, T. Interview self-disclosure as a function of length of modeling and descriptive instructions. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 356–362.

Meichenbaum, D. Examination of model characteristics in reducing avoidance behavior. *Journal of Personality and Social Psychology*, 1971, *17*, 298–307.

Melamed, B., Hawes, R., Heiby, E., & Glick, J. Use of filmed modeling to reduce uncooperative behavior of children during dental treatment. *Journal of Dental Research*, 1975, *54*, 797–801.

Melamed, B., & Siegel, L. Reduction of anxiety in children facing hospitalization and surgery by use of filmed modeling. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 511–521.

Melamed, B., Weinstein, F., Hawes, R., & Katin-Borland, M. Reduction of fear-related dental management problems with use of filmed modeling. *Journal of the American Dental Association*, 1975, *90*, 822–836.

Morris, L. W., Spiegler, M. D., & Liebert, R. M. Effects of a therapeutic modeling film on cognitive and emotional components of anxiety. *Journal of Clinical Psychology*, 1974, *30*, 219–223.

Nietzel, M. T., & Bernstein, D. A. Effects of instructionally mediated demand on the behavioral assessment of assertiveness. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 500.

O'Connor, R. Modification of social withdrawal through symbolic modeling. *Journal of Applied Behavior Analysis*, 1969, *2*, 15–22.

O'Connor, R. Relative efficacy of modeling, shaping, and the combined procedures for modification of social withdrawal. *Journal of Abnormal Psychology*, 1972, *79*, 327–334.

Paul, G. L., & Bernstein, D. A. *Anxiety and clinical problems: Systematic desensitization and related techniques.* Morristown, N.J.: General Learning Press, 1973.

Rathus, S. A. Instigation of assertive behavior through videotape-mediated assertive models and directed practice. *Behaviour Research and Therapy*, 1973, *11*, 57–65.

Reeder, C. W., & Kunce, J. T. Modeling techniques, drug abstinence behavior, and heroin addicts: A pilot study. *Journal of Counseling Psychology*, 1976, *23*, 560–562.

Rich, A. R., & Schroeder, H. E. Research issues in assertiveness training. *Psychological Bulletin*, 1976, *83*, 1081–1096.

Rosenkrans, M. A. Imitation in children as a function of perceived similarity to a social model and vicarious reinforcement. *Journal of Personality and Social Psychology*, 1967, *7*, 307–315.

Sarason, I. Verbal learning, modeling, and juvenile delinquency. *American Psychologist*, 1968, *23*, 254–266.

Schwartz, R. M., & Gottman, J. M. Toward a task analysis of assertive behavior. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 910–920.

Shaw, D. W., & Thoresen, C. E. Effects of modeling and desensitization in reducing dentist phobia. *Journal of Counseling Psychology*, 1974, *21*, 415–420.

Spiegler, M., Liebert, R., McMains, M., & Fernandez, L. Experimental development of a modeling treatment to extinguish persistent avoidance behavior. In R. Rubin & C. Franks (Eds.), *Advances in behavior therapy, 1968*. New York: Academic Press, 1969.

Thelen, M. H., Fry, R. A., Dollinger, S. J., & Paul, S. C. Use of videotaped models to improve the interpersonal adjustment of delinquents. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 492.

Thelen, M. H., Fryrear, J. L., & Rennie, D. L. Delayed imitation of self-reward standards. *Journal of Experimental Research in Personality*, 1971, *5*, 317–322.

Thelen, M. H., Paul, S. C., Dollinger, S. J., & Roberts, M. C. Response uncertainty and imitation: The interactive effects of age and task options. *Journal of Research in Personality*, 1978, *12*, 370–380.

Thelen, M. H., & Rennie, D. L. The effect of vicarious reinforcement on imitation: A review of the literature. In B. A. Maher (Ed.), *Progress in experimental personality research*. New York: Academic Press, 1972.

Thomas, G. M. Using videotaped modeling to increase attending behavior. *Elementary School Guidance and Counseling*, 1974, *9*, 35–40.

VanCamp, J. Modification of adult aggressive behavior by using modeling films (Doctoral dissertation, Fuller Theological Seminary, 1972). *Dissertation Abstracts International*, 1972, *32*, 7327B. (University Microfilms No. 72-15,870)

Vernon, D. T. A. Use of modeling to modify children's responses to a natural, potentially stressful situation. *Journal of Applied Psychology*, 1973, *58*, 351-356.

Vernon, D. T. A. Modeling and birth order in responses to painful stimuli. *Journal of Personality and Social Psychology*, 1974, *29*, 794-799.

Vernon, D. T. A., & Bailey W. The use of motion pictures in the psychological preparation of children for induction of anesthesia. *Anesthesiology*, 1974, *40*, 68-72.

Weissbrod, C., & Bryan, J. Filmed treatment as an effective fear reduction technique. *Journal of Abnormal Child Psychology*, 1973, *1*, 196-201.

Wincze, J. P., & Caird, W. K. The effects of systematic desensitization and video desensitization in the treatment of essential sexual dysfunction in women. *Behavior Therapy*, 1976, *7*, 335-342.

Wroblewski, P. F., Jacob, T., & Rehm, L. P. The contribution of relaxation to symbolic modeling in the modification of dental fears. *Behaviour Research and Therapy*, 1977, *15*, 113-115.

Yussen, S. R., & Levy, V. M., Jr. Effects of warm and neutral models on the attention of observational learners. *Journal of Experimental Child Psychology*, 1975, *20*, 66-72.

Received February 17, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 721–735

Differential Validity of Employment Tests by Race: A Comprehensive Review and Analysis

John E. Hunter
Michigan State University

Frank L. Schmidt
Personnel Research and Development Center
U.S. Office of Personnel Management and
George Washington University

Ronda Hunter
Institute for Research in Teaching
Michigan State University

This study examined 866 black–white employment test validity pairs from 39 studies for evidence of differential validity beyond that which would be expected on the basis of chance plus various statistical artifacts. The data in this study, unlike those in previous studies of differential validity, were free of Type I bias induced by data preselection. The results support the hypothesis that findings of apparent differential validity in samples are produced by the operation of chance and a number of statistical artifacts and indicate that true differential validity probably does not exist.

The question of whether traditional employment tests are appropriate for blacks and other minorities is an important one today. Researchers have approached this question from two different directions: from the point of view of subgroup validity differences and from the point of view of selection fairness. Although these two phenomena are related, they are by no means identical. For example, a test with equal subgroup validities can nevertheless be unfair, under certain circumstances, when used in selection. In earlier articles, we examined the properties of various models of test fairness (Hunter & Schmidt, 1976; Hunter, Schmidt, & Rauschenberger, 1977; Schmidt & Hunter, 1974). For other treatments of selection fairness, see Gross and Su (1975), Novick and Petersen (1976), and Cronbach (1976).

Differential and Single-Group Validity: Definitions and Research

Research in the area of subgroup validity differences has focused on two hypotheses: the single-group validity hypothesis and the differential validity hypothesis. The single-group validity hypothesis states that, for at least some tests, validity in the applicant population is zero for one group (e.g., blacks) but not for the other (e.g., whites), that is, $\rho_1 = 0 < \rho_2$. The research evidence against this hypothesis is now overwhelming (Hunter & Schmidt, 1978). Four different studies based on cumulative available research results (Boehm, 1977; Katzell & Dyer, 1977; O'Connor, Wexley, & Alexander, 1975; Schmidt, Berner, & Hunter, 1973) all found that the frequency of single-group validity does not exceed that which would occur by chance alone given equal population validities.

The differential validity hypothesis, on the other hand, is more general. It states merely that the validities in the two applicant populations are unequal, that is, $\rho_1 \neq \rho_2$. The status of this hypothesis is less certain than that of the single-group validity hypothesis. Because of small sample sizes and other problems, individual studies have limited statistical power to detect differential validity. Studies based on cumulative research results but focusing on the single-group validity hypothesis likewise have limited statistical power to detect differential validity and thus

Requests for reprints should be sent to Frank L. Schmidt, Personnel Research and Development Center, U.S. Office of Personnel Management, 1900 E Street, N.W., Washington, D.C. 20415.

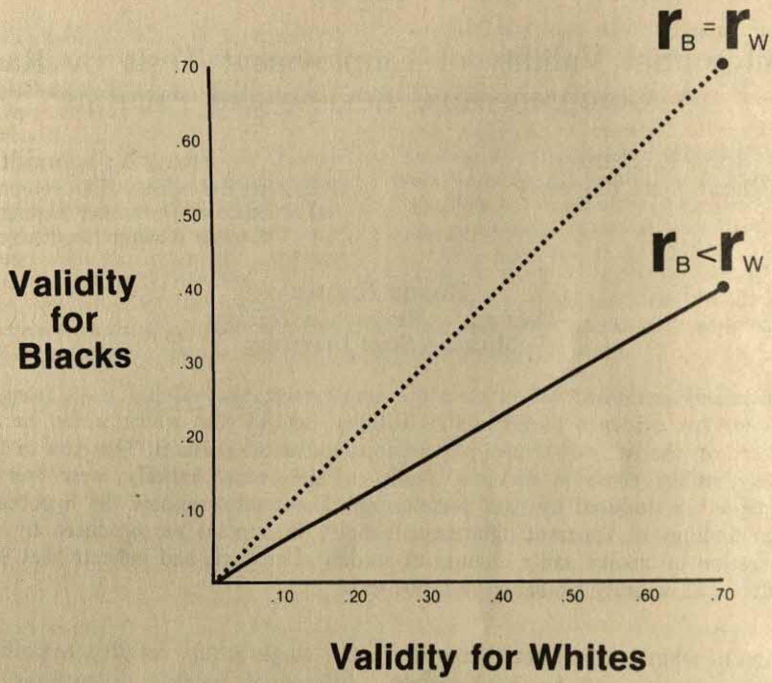

Figure 1. The level-based differential validity hypothesis ($r_B < r_W$) and the null hypothesis ($r_B = r_W$); B = blacks; W = whites.

provide only weak tests of the differential validity hypothesis (Hunter & Schmidt, 1978). Studies based on cumulative research results and focusing specifically on the differential validity hypothesis can have high statistical power for detecting differential validity. To date three such studies have examined the evidence for differential validity (Boehm, 1972; Boehm, 1977; Katzell & Dyer, 1977). However, these researchers preselected the data to be examined in such a way as to induce a massive Type I error bias in their results. Specifically, to be included in their analyses, at least one of the sample validities in a black–white validity pair had to be statistically significant. Hunter and Schmidt (1978) have shown that such preselection of data leads to the appearance of differential validity beyond chance levels even when there is in fact no differential validity in the parent populations. Thus the finding in these three studies that sample differential validity occurred in excess of chance frequency is difficult to interpret. A major purpose of this study is to examine the differential validity hypothesis using data that are not preselected.

It has been suggested (Boehm, 1972; Bray & Moses, 1972) that validity differences by race may be associated with the use of subjective criteria, such as ratings and rankings, and that such differences are infrequent when more objective criteria are employed. This hypothesis has found no support in connection with single-group validity. Schmidt et al. (1973) found that the frequency of single-group validity was at chance levels for both kinds of criteria. (The proportion of validity pairs showing single-group validity was significantly higher for subjective than for objective criterion measures [.37 vs. .20, $p < .001$], but this difference was predicted and fully explained by differences between the two data sets in individual sample sizes, differences between black and white sample sizes, and general level of test validity.) The second purpose of the present research is to test this hypothesis in connection with differential validity.

A third purpose of this research is to examine the differential validity hypothesis separately for different levels of validity. It may be that tests of generally low validity show little or no differential validity, while tests of higher validity show substantial differential

validity. This hypothesis is depicted graphically in Figure 1. This figure shows the predicted graph of test validity for blacks as a function of validity for whites. If a test is invalid for whites, it is predicted to be invalid for blacks also, and hence the function begins at the point (0, 0). The validity for blacks is predicted to increase as the validity for whites increases, but at a slower rate. Hence the function shown in Figure 1 is always less than the broken line, $r_B = r_W$ (B = black; W = white).

Although Figure 1 does illustrate the hypothesis that degree of differential validity is a function of level of test validity, the figure can be criticized in two respects. First, the comparison of the validity for whites and for blacks is made in the form of a comparison of the observed curve with the hypothetical broken line, $r_B = r_W$, rather than in the form of a direct comparison of curves for each group separately. Second, and more crucial, a plot of actual data (as opposed to population values) in the form of Figure 1 would show a definite bias produced by the sampling error in the correlations for whites. The sampling error in the white validity coefficients would function as an error of measurement in the independent variable in that it would result in a false reduction of the slope of the regression line; that is, a plot of real data would be biased in the direction of the differential validity hypothesis even if that hypothesis were false. This follows from the fact that measurement error (unreliability) in the independent variable—but not measurement error in the dependent variable—acts to attenuate the slope (cf. Hunter & Schmidt, 1976, for further development of this point).

An alternative manner of plotting the same hypothesis is shown in Figure 2. In Figure 2 the basic validity of the test is not defined as the validity for whites but as the average of the validity for whites and the validity for blacks. The separate validities for each group are then plotted as a function of the within-pair average, producing separate functions for whites and blacks. These two functions can then be compared in the usual way. First, if validity is zero for either group, then it is expected to be zero for the other group, that is, both functions start out at (0, 0). As the average validity for the test increases, the validity for each group increases. However, the validity for whites is expected to increase more rapidly than the validity for blacks, and

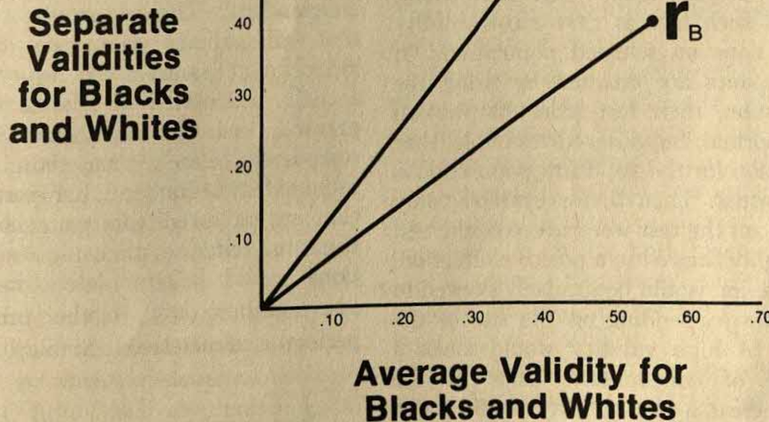

Figure 2. An alternate expression of the level-based differential validity hypothesis. (B = blacks; W = whites.)

hence the hypothesized curve for whites always lies above the hypothesized curve for blacks.

Data Analysis: Methods and Problems

A special form of chi-square test was developed for use in testing the differential validity hypothesis: The coefficients in each validity pair (r_B and r_W) are first transformed to Fisher's zs, that is, to z_B and z_W, respectively. For each validity pair, the chi-square is then as follows:

$$\chi^2 = \frac{(z_W - z_B)^2}{\dfrac{1}{N_W - 3} + \dfrac{1}{N_B - 3}},$$

where N_W is the sample size for whites and N_B is the sample size for blacks. Note that the right side of this equation is merely a squared Z score, as required by the definition of chi-square. The Z scores result, of course, from standardization of the validity differences after conversion of validities to Fisher's z form. For a single validity pair, this chi-square has only one degree of freedom, but (unlike a significance test between correlations) it can be cumulated over all pairs in a given range of average validity or over all pairs across all levels of validity to provide a test with high statistical power. However, because of the nature of the data that must be used in any study of differential validity findings from the literature, this chi-square test (and any other test) is characterized by a number of sources of Type I error bias.

Effects of nonnormality. The first factor that creates such bias is that most validity studies are done on selected populations. If such participants are selected by using the test in question, their test score distribution will be nonnormal. Suppose, for example, that the distribution for the applicant population is bivariate normal. Then if, for example, only the top 50% on the test were selected, the test would be highly skewed in a positive direction, and the criterion would be similarly skewed to a lesser extent, depending on the size of the correlation. (A high validity would mean a high degree of concomitant skew in the criterion, whereas a low validity would mean that the criterion would be approximately normal.) We found no literature examining the impact of these nonnormal distributions on the standard error of the correlation coefficient. However, we were able to derive a fairly good approximation. Our calculations show, for example, that for a 50% selection ratio and a validity of .5 in the applicant population (reasonable assumptions; see Schmidt & Hunter, 1977; Schmidt, Hunter, & Urry, 1976), the standard error of the selected sample validity coefficient is 8% larger than the full population standard error derived from the assumption of normality. The probability of finding a false reading of differential validity in such a situation is not the usual 5%, but rather 7%. (These calculations are available from the senior author.[1]) If such studies were added into a cumulative chi-square test, then each such study would add an expected increment of 1.16 to the chi-square, instead of the 1.00 that would be added if the full applicant population were sampled. This results, of course, in a Type I bias. In this study, we deal with a total of 866 black–white validity pairs. In the absence of Type I bias, the expected value of the cumulative chi-square over these validity pairs would be its degrees of freedom, 866. In the presence of this Type I bias, the expected value of the chi-square under the null hypothesis (that is, given the complete absence of any true differential validity) would be 1.16 (866), or 1,004.6 in our example, which would be deemed significant beyond the .0001 level.

Violations of independence. The second factor creating Type I error bias is the fact that not all validity pairs are statistically independent. The chi-square test assumes that all validity pairs are computed on independent samples. If several pairs are actually calculated on the same sample of persons (as is the case in nearly every study considered below), then those pairs are to some extent redundant. For example, suppose that several predictors were correlated with the same criterion; then the resulting correlations would be completely independent of one another only if the predictors were perfectly uncorrelated. Similarly, if one pre-

[1] These calculations are quite lengthy and are therefore not included here. They will, however, be presented in a future article.

dictor were correlated with several criterion measures, then the resulting correlations would be independent of one another only if the criterion measures were uncorrelated with each other. The effect of this partial redundancy produced by nonzero correlations between multiple predictors or between multiple criteria is difficult to calculate, and no attempt is made to correct for it below. Instead we simply state an example by way of warning. Suppose we had run a chi-square statistic over 100 validity pairs and had obtained an observed value of 128. Then, had those pairs been independent, we would have used the known facts about the chi-square distribution to conclude that the chance expected value of the chi-square statistic would be 100 and the standard deviation would be the square root of twice the degrees of freedom, that is, $\sqrt{(200)} = 14.14$, and hence an observed value of 128 would be significant at the 5% level. (Recall that a chi-square distribution with 100 degrees of freedom is essentially normal.) However, suppose we were told that an error had been made and that one of the numbers had been inadvertently written down 10 times. Then the expected value of the "chi-square" statistic would still be the nominal degrees of freedom, 100. To obtain the variance, we would note that this statistic was the sum of 91 independent entries, 90 of which had the usual chi-square variance of 2, whereas the last entry was repeated 10 times and hence had a variance of $10 \times 10 \times 2$, or 200. Thus, the true standard deviation would be $\sqrt{(180 + 200)} = 19.49$, and an observed value of 128 would no longer be statistically significant. Obviously, use of the smaller, erroneous standard deviation leads to a Type I error bias in the statistical test.

The typical correlation between alternate predictors and alternate criteria in the studies considered here is far less than 1.00. Thus the Type I error bias is probably not as great as in this hypothetical example. However, there is a Type I error bias. If in fact there is no true differential validity, the probability of rejecting this true null hypothesis is greater than .05; that is, when independence assumptions are not fully met, the nominal .05 alpha level is not the operational alpha level. The true alpha level is numerically larger, which biases the statistical test in favor of the alternate hypothesis that there is true differential validity.

When lack of independence is incomplete—which is usually the case—it is typically not possible to compute the correct standard deviation or degrees of freedom. If it were possible, the resultant chi-square statistic would of course not have a Type I error bias. However, this chi-square would have somewhat less statistical power than would be present if there were no violations of independence to begin with and the usual chi-square formulas were used to compute the standard deviation and the degrees of freedom.

Differential range restriction. The above two factors tend to produce the false appearance of differential validity by violating assumptions basic to the chi-square test. One other factor—differential range restriction by race—also tends to produce artifactual differential validity, although it does not violate strictly statistical assumptions. Such differential validity is artifactual in that it is produced by purely statistical factors that have nothing to do with substantive black–white cultural differences; that is, such apparent differential validity will manifest itself even when the validities of interest—the applicant population validities—are identical for blacks and whites.

The effect of range restriction on test validity was examined in detail by Schmidt et al. (1976). Suppose that the correlation in the applicant population is .50 but that only the top 40% of the applicants were selected and that the validity study was done on this group. The expected validity on the selected applicants would be only .31.

The situation is fundamentally more complicated when the validities of two groups are to be compared, because one must then consider the possibility that selection does not act equally on the two groups. Suppose, for example, that the validity in the applicant population is .50 for both blacks and whites; that is, suppose that there is no true differential validity. Suppose further that the cutoff score is set such that 40% of the whites are selected. The validity coefficient for the selected whites would be lowered to .31 by this artifactual restriction in range. Would the same happen to blacks? That depends on the mean black

test score. If the mean for blacks were the same as the mean for whites, then the effect of restriction in range would be the same. But this is typically not the case. Rather, there is usually a difference of about 1 standard deviation between whites and blacks on the test. If blacks have a test mean that is 1 standard deviation below the mean for whites, then the cutoff score is not .25 standard deviations above the black mean, as it is for whites; rather, it is 1.25 standard deviations above the black mean. Thus the selection ratio for blacks is not 40%, but 11%. Since blacks are much more severely selected, their test scores are much more severely restricted in range. In fact, the expected black validity is not .31, but .23; that is, selection at the same cutoff score for whites and blacks results in differential restriction in range and hence in an artifactual difference in the observed validities in the selected group.[2] The information necessary to correct validities for differential range restriction is rarely presented in studies.

Would an artifactual difference on the order of .30 versus .23 be significant in existing differential validity data? In the data reviewed for this study, the average Ns per validity pair were 140 for whites and 75 for blacks. If one considers a single study with these Ns, then this difference adds .231 to the chi-square (with 1 degree of freedom). This addition is nowhere near significant. Thus, this sort of artifact is not immediately detectable in single studies. But suppose that 866 such studies were run, as in this study. Then the chi-square pooled over the 866 validity pairs would be inflated beyond sampling error by 866 (.231) = 200; that is, the expected value of the chi-square would be 866 + 200 = 1,066. This would raise the probability of obtaining a significant cumulative chi-square from the 5% chance level prevailing when studies are done with applicant populations to the 99% prevailing when calculations are based on the selected applicants. [This calculation takes into account the fact that a noncentral chi-square has a larger standard deviation. For the central chi-square that represents the null hypothesis, the expected value would be its degrees of freedom, 866, and the standard deviation would be $[2(866)]^{\frac{1}{2}}$, or 41.64. For the noncentral chi-square, which represents the true state of nature here, the expected value would be 866 + .231 (866) = 1,066, and the standard deviation would be $\{[2+4(.231)]\, 866\}^{\frac{1}{2}} = 50.32$.] Had 866 separate chi-square tests been done, this same fact would have taken a different form. If the studies had been conducted on the applicant populations, then about 5% of the studies would have produced a false reading of significant differential validity. But in the selected applicant population, the probability of falsely detecting differential validity (i.e., the probability of a significant chi-square) is 7.5%. Over a large number of studies, this would lead one to falsely conclude that differential validity would be found in 2.5% (i.e., 7.5% minus 5.0%) of the corresponding job–test combinations.

Statistical power. In contrast to the situation prevailing with respect to Type I error biases, Type II errors are very well controlled in these data. Because the data set contains a large number of validity pairs (866) based on a much larger number of data points (see Discussion section), statistical power is very high. The probability of failing to detect true differential validity is quite low. Recall from our discussion of differential range restriction that the power to detect the small 7-point correlation difference between .23 and .30 is .99. Even for the trivial difference of .05 (Schmidt & Hunter, 1978), statistical power is .77.

In summary, the chi-square test used in this study was affected by three sources of Type I error. Obviously, the usual .05 level of significance was not appropriate for use with these data. In the case of the individual chi-square tests, for example, the expected percentage of tests significant under the null hypothesis (no true differential validity), using $\alpha = .05$, was not 5% but some substantially larger number. Based on the discussion above, it is our judgment that a conservative criterion would hold that findings of 10% or less disconfirm the differential validity hypoth-

---

[2] Even in unselected groups, tests scores of blacks have sometimes been found to be less variable (e.g., see Shuey, 1958). However, these findings have been based on school populations and may not apply to applicant populations.

esis and that proportions in the 5%–7% range constitute compelling evidence against the differential validity hypothesis.

A second test. In addition to the chi-square test, differences in mean validities between blacks and whites, within validity intervals and across all levels of validity, were examined using a critical ratio (*CR*) test. In this test, the error variance terms used were computed directly rather than being based on an assumed theoretical error distribution, as in the case of our chi-square test. This test can be represented as follows:

$$CR = \frac{E(r_W - r_B)}{SD_d/\sqrt{n-1}},$$

where the numerator is the average difference between the black and white validities, SD_d is the standard deviation of these differences, and n is the number of validity pairs. The advantage of this test is that departures from normality in test score or criterion distributions do not create Type I biases. The test is, however, affected by departures from independence and by differential range restriction. These effects operate to produce Type I biases of the same kind produced in our chi-square test, that is, the true alpha level will be numerically larger than the nominal alpha level.

Method

A careful review of published studies produced 39 studies reporting employment-test validities separately by race. These studies contained a total of 866 pairs of validity coefficients for which sample sizes were reported. A total of 532, or 61%, of the validity pairs were based on subjective criterion measures; 334, or 39%, were computed using objective criterion measures. These figures are shown by individual study in Table 1. All ratings, rankings, and so on, and grades in training (when not based on performance measures) were considered subjective criteria; performance measures such as quality and quantity of output, job sample measures of proficiency, errors, attendance, tenure, written job knowledge tests, and the like were considered objective criteria.

All coefficients were converted to Fisher's z, and values for the chi-square test described earlier were computed for each validity pair. The percentage of chi-square values reaching significance was calculated for (a) the overall group of 866 pairs, (b) each validity interval examined, and (c) the objective and subjective criterion measures. These chi-square values were also cumulated to produce overall values within each of these data categories. Critical ratios for level differences between blacks and whites were also run within each of these categories.

Results

The proportion of chi-square values significant at the .05 level in the total sample of 866 validity pairs was .09, which is less than our stipulated criterion of .10. The cumulative chi-square for the 866 pairs was 1,123.02, which is not significantly different from the value that might be produced artifactually given our earlier example of differential range restriction (1,066). Nonnormality of test and criterion distributions, induced by selection on the test, and violations of the independence assumption operated, of course, along with differential range restriction to bias the obtained chi-square of 1,123 in a Type I direction. Because of the large sample sizes, both these figures are nominally significant beyond the .01 level, but, as shown earlier, they are well within the range of results to be expected from Type I biases alone in the absence of any true differential validity. The overall mean racial difference in validity was .02 (whites higher). This difference, trivial from a practical point of view, is nominally statistically significant.

Figure 3 shows the results relevant to the hypothesis, which is illustrated in Figure 2, that degree of differential validity increases with level of validity. As the reader may recall from the discussion of Figure 2, the black and white validities (y-axis) are plotted separately as a function of the average of the two validities (x-axis). For positive validity, the curve is quite straightforward: There is no apparent evidence for differential validity. The only noticeable difference in the two curves is found in one point (the highest point) based on only 11 validity pairs, and for this point the mean black validity is higher than the mean white validity. On the other hand, the data for negative validities are strange and unexpected. All variables were either scored such that expected validity would be positive or were reflected to achieve the same effect. True negative validities were thus highly unlikely. Therefore, the negative correlations for both groups should represent sampling error, that is, negative deviations

Table 1
Distribution of Validity Pairs Across Studies

Study	Criteria		
	Subjective	Objective	Total
Campbell, Crooks, Mahoney, & Rock (Note 1)	92	46	138
Campbell, Pike, Flaugher, & Mahoney (Note 2)	18	0	18
Campbell, Pike, & Flaugher (Note 3)	0	8	8
Campion & Freihoff (Note 4)	0	10	10
Farr (Note 5)			
Study 1	5	0	5
Study 2	0	90	90
Farr, O'Leary, & Bartlett (1971)			
Study 1	2	19	21
Study 2	52	8	60
Farr, O'Leary, Pfeiffer, Goldstein, & Bartlett (Note 6)	46	0	46
Flaugher, Campbell, & Pike (Note 7)	36	0	36
Fox & Lefkowitz (1974)	18	9	27
Gael & Grant (1972)	0	35	35
Gael, Grant, & Ritchie (1975a)	0	11	11
Gael, Grant, & Ritchie (1975b)	0	11	11
Grant & Bray (1970)	0	8	8
Kirkpatrick, Ewen, Barrett, & Katzell (1968)			
Study 1	4	20	24
Study 2	28	0	28
Study 3[a]	0	6	6
Study 5	8	8	16
Lefkowitz (1972)	0	8	8
Lopez (1966)	4	12	16
Mitchell, Albright, & McMurry (1968)	1	1	2
O'Leary, Farr, & Bartlett (Note 8)[b]	121	0	121
Ruda & Albright (1968)	0	2	2
Tenopyr (Note 9)	36	0	36
Toole, Gavin, Murdy, & Sells (1972)	30	0	30
U.S. Department of Labor (Note 10)[c]	9	0	9
U.S. Department of Labor (Note 11)	1	0	1
U.S. Department of Labor (Note 12)	1	0	1
U.S. Department of Labor (Note 13)	1	0	1
U.S. Department of Labor (Note 14)	1	0	1
U.S. Department of Labor (Note 15)	1	0	1
U.S. Department of Labor (Note 16)	1	0	1
U.S. Department of Labor (Note 17)	1	0	1
U.S. Department of Labor (Note 18)	1	0	1
U.S. Department of Labor (Note 19)	1	0	1
U.S. Department of Labor (Note 20)	1	0	1
Wollowick, Greenwood, & McNamara (1969)	12	12	24
Wood (Note 21)	0	10	10
Total	532	334	866

[a] Included data on Spanish Americans, which were excluded for purposes of this analysis.
[b] Includes all data not published in Farr, O'Leary, & Bartlett (1971).
[c] Includes one American Indian.

from a population in which the correlation is close to zero. If this were true, then we would expect no differences between the curves for the negative validity samples. Thus, the observed difference (which is significant) suggests that something other than routine sampling error occurred in the negative validity samples.

The basic difference in the results for positive and negative validity samples was brought out by a second analysis: a count of the number of validity pairs that were significantly

different, that is, showed significant chi-square values. In the positive validity class, 45 out of 712 pairs were significantly different, that is, 6%, which is near even the nominal chance level. However, for the negative validity class, 32 out of 154 pairs were significantly different, that is 21%, which is considerably beyond the nominal chance level and perhaps even beyond the true chance level for these data.

A search of the data showed that the unexpected negative validities were concentrated in three studies: Farr, O'Leary, Pfeiffer, Goldstein, and Bartlett (Note 6), O'Leary, Farr, and Bartlett (Note 8), and Toole, Gavin, Murdy, and Sells (1972). The Farr et al. study contained 46 pairs of correlations; 23 of these were based on a sample of 31 black and 95 white clerical workers. In the group of 95 whites, only 5 of the validity coefficients were negative, and these were all near 0, ranging from $-.05$ to $-.10$. In the group of 31 blacks, all 23 coefficients were negative, and the average value was $-.24$. In this same study, another 23 validity pairs were based on a sample of 51 black and 158 white insurance workers. In the group of 158 whites, none of the validity coefficients was negative. In the group of 51 blacks, all but one coefficient was negative, and its value was .01. The average of the coefficients was $-.15$. In all cases, these validity coefficients were for types of tests that are valid in a positive direction in other black samples.

In one substudy of the O'Leary et al. study, 48 pairs of correlations were presented; 24 of these were based on a sample of 31 black and 60 white clerical machine operators. In the group of 60 whites, only 4 coefficients were negative, and all of these were near 0, ranging from $-.02$ to $-.07$. In the group of 31 blacks, 18 of the 24 coefficients were negative, and the average value was $-.21$. In this same study, another 24 validity pairs were based on a sample of 24 black and 106 white "miscellaneous" clerical workers. In the group of 106 whites, only 1 coefficient was

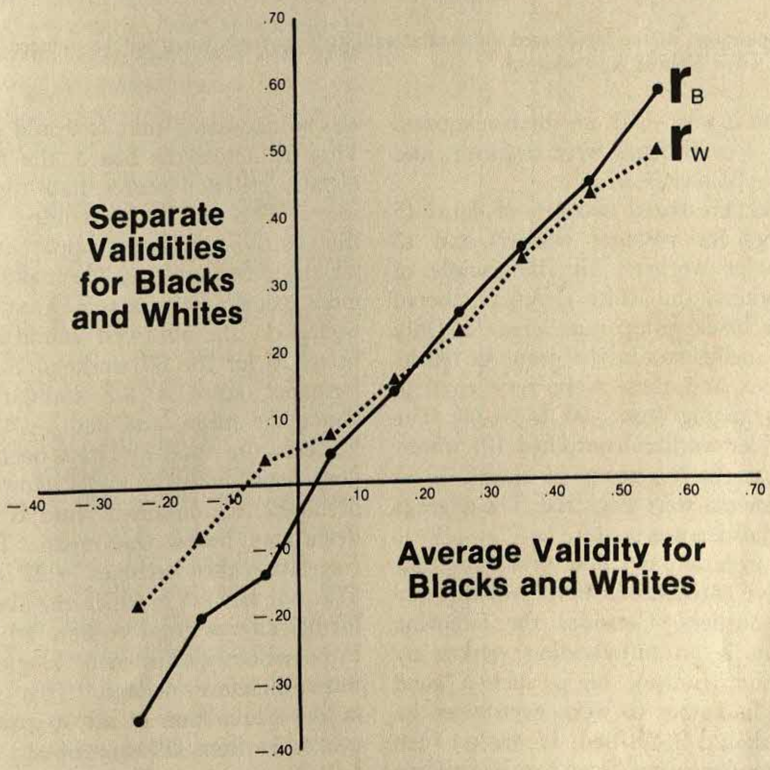

Figure 3. Graphic test of the level-based differential validity hypothesis based on all 866 validity pairs. (B = blacks; W = whites.)

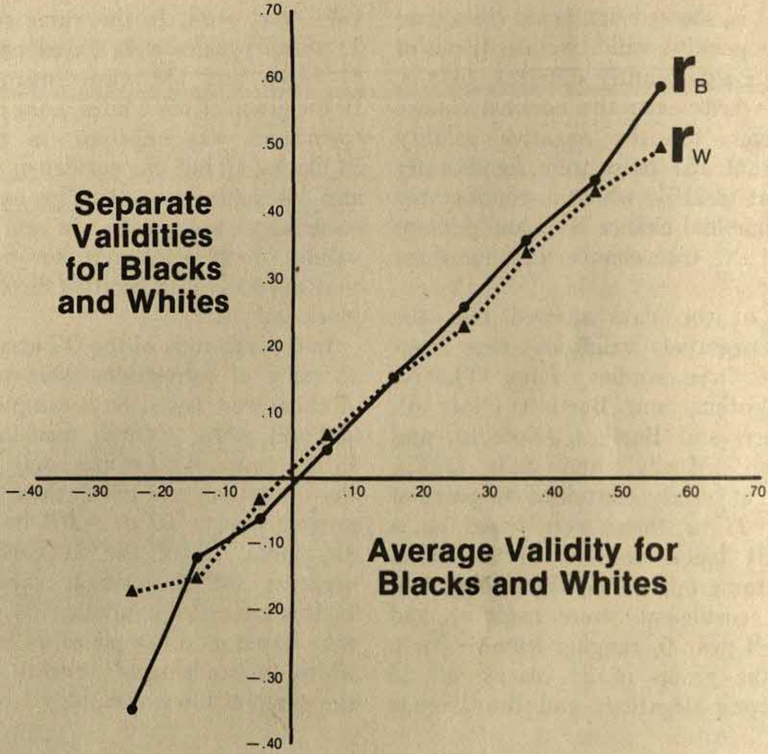

Figure 4. Graphic test of the level-based differential validity hypothesis based on the reduced data set of 781 validity pairs. (B = blacks; W = whites.)

negative, and it was −.01. In the black group of 24, only 2 coefficients were negative, and these were −.01 and −.07.

Toole et al. presented two sets of data: 15 validity pairs for younger workers and 15 pairs for older workers. In the sample of younger workers, the white group numbered 288 and the black group numbered 75. Only 4 of the 15 coefficients in the group of blacks were negative, and these were very small in magnitude, ranging from .00 to −.08. The sample of older workers contained 121 whites and 36 blacks. In the group of blacks, 13 of the 15 coefficients were negative. The average of the 15 validities was −.17.

Our interpretation of these results is that these negative correlations were probably the product of outliers. Consider the following parable. Sam is an outstanding worker by every criterion. Indeed, he is such a good worker that he comes to work even when he is sick and should be in bed. It was on such a day that the tests were given for the validity study. Sam did his best on the tests, but he was so nauseated that he could hardly think. Thus all across the board, the man with the highest criterion scores had the lowest test scores. This one outlier tended to cancel out the positive correlation apparent in the data for the other subjects. To make the example more specific, let us assume that had Sam not been sick, the observed validity would have been .20 for the 30 workers. Sam's job performance score is 2.5 standard deviations above the mean, and had he not been sick, his test score would also have been 2.5 standard deviations above the mean. However, because of illness, his obtained score is 2.5 standard deviations below the mean. The observed correlation then becomes −.22 instead of .20. The four samples in which the average validity for blacks was negative were all small enough to be seriously affected by single outliers and were therefore deleted. This step resulted in the elimination of all 46 pairs from Farr et al., 24 from O'Leary et al., and 15 pairs from Toole et al. In total, 85 validity pairs were dropped.

Figure 4 shows the comparison of black and white validity coefficients for the 781 pairs of validity coefficients remaining after the suspect data were deleted. The only point at which the curves differ noticeably is the highest point, which is based on only 11 pairs, and at this point the mean validity is higher for blacks than for whites. This figure clearly disconfirms the hypothesis that differential validity increases with validity level. Moreover, the two curves are now essentially equal for the negative validities resulting from sampling error, just as they are equal for the studies with positive validity.

The proportion of chi-square values significant at the .05 level was .067 in the reduced data set of 781 pairs. With 781 cases, the difference between 6.7% and 5% is nominally significant ($p < .05$, binominal test). But this small deviation from chance is in fact less than would be expected from the combined effects of the three known sources of Type I bias. The cumulative chi-square value was 874.4, which is nominally statistically significant ($p < .01$). But, again, the number of cases (781) is very large, and the result is actually miniscule when compared with that predicted from selection-induced skew in score distribution, violations of independence, and differential range restriction. In fact, the expected value of this chi-square in our earlier example of the range restriction artifact was 1.23 (781) = 961. The other two sources of Type I bias discussed earlier, of course, also operated here. The overall difference between the mean validities for whites and blacks was zero (to two digits), and the corresponding critical ratio was nowhere near significance ($CR = -.25$, $p < .60$). Thus these tests, taken together, suggest the absence of true differential validity in these data.

Table 2 shows the results of the three statistical tests separately for each level of average test validity. For example, in the validity category .51–.60, the mean validity for blacks is .58 and the mean validity for whites is .49. This mean difference is significant ($CR = 3.33$, $p < .002$) and leads to a significant chi-square, $\chi^2(11) = 23.80$. However, a check of the number of cases shows that for this validity interval, 2 of the 11 pairs are significantly different; a number this small could

Table 2
Differential Validity Findings by Level of Validity[a]

Item	Validity levels								
	−.30 to −.20	−.19 to −.10	−.09 to .00	.01 to .10	.11 to .20	.21 to .30	.31 to .40	.41 to .50	.51 to .60
No. pairs	3	22	59	184	206	186	80	30	11
No. significant	0	2	8	7	11	11	5	5	2
Proportion significant	.00	.09	.14	.04	.05	.06	.06	.17	.18
Cumulative χ^2	5.1, *ns*	35.4*	76.9*	174.8, *ns*	211.7, *ns*	223.6*	83.0, *ns*	40.0, *ns*	23.8*
M black validity	−.35	−.12	−.06	.04	.15	.26	.35	.44	.58
M white validity	−.18	−.16	−.03	.06	.15	.23	.34	.43	.49
Black *SD*	.12	.14	.13	.10	.10	.09	.08	.08	.06
White *SD*	.15	.12	.13	.10	.10	.09	.08	.08	.05
Critical ratio (level differences)	1.11	−.70	.82	1.27	.29	−2.58*	−.90	−.44	−3.33*

[a] Reduced data set; number of validity pairs = 781.
* $p < .05$ in absence of Type I biases; see text for discussion of these biases.

Table 3
Comparison of Differential Validity Findings for Objective and Subjective Criterion Measures

Item	Objective criterion[a]	Subjective criterion: All data	Subjective criterion: Reduced set[b]
No. pairs	334	532	447
Proportion of χ^2s significant	.08*	.10*	.06
Cumulative χ^2	399.3*	723.7*	475.2
M racial difference in validity[c]	−.01	−.04*	.01

[a] All data; none of the validity pairs dropped involved objective criteria.
[b] Data deleted as described in text.
[c] Negative values indicate higher white validities.
* $p < .05$ in absence of Type I biases; see text for discussion of these biases.

be due to chance. Evidence in this direction is found in the adjacent column for the validity interval .41–.50; here, despite containing nearly three times as many cases, the critical ratio is not significant.

The close comparability of mean black and white validities across the average validity intervals is particularly striking.[3] This fact is reflected in the values of the critical ratio assessing level differences in each of the validity intervals (last row of Table 2). Of the nine critical ratios, only two were statistically significant. One was the previously discussed validity interval .51–.60, which was based on only 11 validity pairs. In the other interval showing significance (.21–.30), the mean validity difference between blacks and whites (.26 − .23 = .03) was trivial in magnitude and was not replicated in the intervals on either side.

The proportion of validity pairs in which the difference was significant varied across the intervals from .00 to .18; the unweighted average was .087. Four of the nine cumulative chi-square statistics were nominally significant ($p < .05$).

In interpreting all the statistics in Table 2, one must again bear in mind the Type I biases that operated. In light of these biases, the statistics for validity intervals, like the whole-sample statistics discussed above, strongly suggest the absence of any true (population) differential validity.

Table 3 summarizes differential validity findings separately for objective and subjective criterion measures. These findings provide no evidence to support the hypothesis that differential validity is associated with subjective criterion measures. The overall proportion of significant chi-squares for subjective criteria was .10 in the complete data set and .06 in the reduced set. For objective criteria, the proportion was .08 in both data sets. Given the known sources of Type I bias in these data, all proportions must be considered in the chance range. Interestingly, the only cumulative chi-square that did not reach significance pertained to subjective criteria (reduced data set). The racial difference in mean validities appears to be negligible for both criterion types. To conserve space, the graphs of mean black and white validity values against average validity intervals are not shown separately for subjective and objective criteria. Both graphs for objective criteria are essentially identical to Figure 4. In the case of subjective criteria, the graph based on all data is very similar to Figure 3, and the reduced data set produces a graph essentially identical to Figure 4. Thus, the evidence appears to indicate that our overall conclusion about the nonviability of the differential validity hypothesis can safely be generalized across criterion types.

Discussion

The complete set of 866 validity pairs in this study reflects 185,487 data points (120,294 for whites and 65,193 for blacks). The reduced set of 781 validity pairs is based on 173,190 data points (111,333 for whites and 61,857 for blacks). Thus, statistical power in these

[3] The table computed on the unreduced set of 866 validity pairs is available from the authors.

analyses is almost infinite in comparison with usual research standards; that is, although Type I error biases are a problem in these data, there are no problems of Type II error biases. The probability of a Type II error, given even trivial validity differences, is close to zero. Yet no differences were found beyond those expected on the basis of Type I biases plus chance. In fact, the differences found were smaller than one might expect to find on the basis of the combined effect of these two factors. These results appear to clearly disconfirm the differential validity hypothesis. The conclusion must be that tests which rank order whites successfully with respect to some given criterion also rank order blacks equally successfully.

An argument occasionally heard against the earlier findings that single-group validity does not exceed chance frequencies may be offered in connection with the findings of this study. This argument can be summarized as follows. The frequency of differential validity does not exceed chance levels in the pooled data as a whole, and this finding does indicate that true differential validity is quite uncommon. But it may nevertheless be that specific instances of statistically significant validity differences observed in samples may reflect true differential validity in particular applicant populations. This argument must be addressed in terms of the overwhelming evidence. Recall that statistical power in this and similar studies is extremely high. If true differential validity existed in more than a tiny fraction of the applicant populations included in the study, it would be detected in the analysis of the data as a whole. Yet no such effect was observed in the data. The only possible conclusion, then, is that the probability is vanishingly small that any specific validity difference observed in samples reflects a true applicant population difference.

Some may object to the analyses performed on the reduced data set. In our judgment, the elimination of the 85 small-sample validity pairs (9.8% of all pairs) from the negative validity data set has been amply justified, but those who may disagree must still account for the fact that in the unreduced data set, the frequency of differential validity was clearly in the chance range (6%), even ignoring Type I biases, in the entire range of positive validities. Recall that all predictors were either scored such that expected validity was positive or were reflected to achieve the same effect. True negative validities (e.g., a negative population correlation between a cognitive ability and job performance), if they exist, must be regarded as anomalies. Clearly, most psychologists are essentially interested only in the positive validity range, and here the results for the complete and reduced data sets agree in strongly indicating the absence of any true differential validity.

What are the implications of these findings for the question of the appropriateness of personnel tests for blacks? As indicated earlier, lack of differential validity does not assure test fairness. For example, under the regression definition of selection fairness, regression line intercepts can be unequal even given equal validities and equal test and criterion variances. According to the regression definition of test fairness, such unequal intercepts denote unfairness to the group with the higher intercept when test scores are used in selection in the same regression equation for both groups. On the other hand, equal validities imply the impossibility of all of the forms of test unfairness specifically stemming from validity differences by race. For example, one potential cause of differences in regression slopes is differences in validity.

A more important point from a theoretical point of view, however, is that most substantive theories or hypotheses of test bias make predictions of differential validity (in addition to predictions of, for example, lower test means for blacks); that is, hypotheses of test bias have been based on the assumption that the actual content of tests, having been based on the content of white middle-class culture, does not mean the same thing psychologically to blacks as it does to whites. If this is in fact true, one result must be differential validity. Consider the following example. Suppose a certain perceptual speed test requires the examinee to follow complicated written directions. If there are differences in English dialect so strong that blacks must employ an internal translation process, two effects will result. First, blacks at a given level of ability in perceptual speed will have a more difficult

task than whites at the same ability levels and will make more errors, leading to lower test scores. Second, and more important, the factor composition of the test will not be the same for the two races. Whereas the test would assess only individual differences in perceptual speed for whites, it would, in addition, measure individual differences in internal translation abilities for blacks. The psychological meaning of the test scores will not be the same for blacks and whites, and thus the pattern of validities the test shows for various external criteria will differ by race. Thus, our failure to find any support for the phenomenon of differential validity in this study constitutes strong evidence against substantive hypotheses of test bias based on the assumption that the meaning of test content differs by race.

Reference Notes

1. Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. *An investigation of sources of bias in the prediction of job performance: A six-year study* (Final Project Report No. PR-73-37). Princeton, N. J.: Educational Testing Service, 1973.
2. Campbell, J. T., Pike, L. W., Flaugher, R. L., & Mahoney, M. H. *Prediction of job performance for Negro and white medical technicians: The prediction of supervisors' ratings from aptitude tests, using a cross-ethnic cross-validation procedure* (Report No. PR-70-18). Princeton, N. J.: Educational Testing Service, 1970.
3. Campbell, J. T., Pike, L. W., & Flaugher, R. L. *Prediction of job performance for Negro and white medical technicians—A regression analysis of potential test bias: Predicting job knowledge from an aptitude battery* (Report No. PR-69-6). Princeton, N. J.: Educational Testing Service, 1969.
4. Campion, J. E., & Freihoff, E. C. Unintentional bias when using racially mixed employee samples for test validation. *Experimental Publication System*, October 1970, *8*, Ms. No. 285-2.
5. Farr, J. L. *The use of work sample and culture-fair tests in the prediction of job success with racially mixed groups.* Paper presented at the meeting of the American Psychological Association, Washington, D. C., September 1971.
6. Farr, J. L., O'Leary, B. S., Pfeiffer, C. M., Goldstein, F. L., & Bartlett, C. J. *Ethnic group membership as a moderator in the prediction of job performance: An examination of some less traditional procedures* (Tech. Rep. No. 2). Washington, D. C.: American Institutes for Research, September 1971.
7. Flaugher, R. L., Campbell, J. T., & Pike, L. W. Ethnic group membership as a moderator of supervisors' ratings (Research Bulletin PR-69-5). Princeton, N. J.: Educational Testing Service, April 1969.
8. O'Leary, B. S., Farr, J. L., & Bartlett, C. J. *Ethnic group membership as a moderator of job performance* (Tech. Rep. No. 1). Washington, D. C.: American Institutes for Research, April 1970.
9. Tenopyr, M. L. *Race and socio-economic status as moderators in predicting machine-shop training success.* Paper presented at the meeting of the American Psychological Association, Washington, D. C., September 1967.
10. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for welder, production line (welding) 812.884* (Tech. Rep. S-447). Washington, D. C.: Author, November 1969.
11. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for ward clerk (medical ser.) 219.388* (Tech. Rep. S-239R74). Washington, D. C.: Author, 1975.
12. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for teller (banking) 212.368* (Tech. Rep. S-259R75). Washington, D. C.: Author, 1975.
13. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for teacher aide, elementary school (education) 099.368* (Tech. Rep. S-398R74). Washington, D. C.: Author, 1974.
14. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for clerk, general office (clerical) 219.388* (Tech. Rep. S-329R74). Washington, D. C.: Author, 1974.
15. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for nurse, licensed practical (medical ser.) 079.378* (Tech. Rep.). Washington, D. C.: Author, 1974.
16. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for keypunch operator (clerical) 213.582* (Tech. Rep. S-180R74). Washington, D. C.: Author, 1974.
17. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for electronics assembler (electronics) 726.781* (Tech. Rep. S-310R74). Washington, D. C.: Author, 1974.
18. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for drafter (civil, 005.281; geological, 010.281; mechanical, 007.281; structural, 005.281)* (Tech. Rep. S-266R74). Washington, D. C.: Author, 1974.
19. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for fork-lift truck operator (any ind.) 922.883* (Tech. Rep. S-131R74). Washington, D. C.: Author, 1974.
20. U. S. Department of Labor, Manpower Administration. *Development of USES aptitude test battery for production mechanic, tin cans (tin ware), 619.380, preventive maintenance mechanic, cup forming (paper goods), 649.380* (Tech. Rep. S-370R75). Washington, D. C.: Author, 1975.

21. Wood, M. T. *Validation of a selection test against a turnover criterion for racial and sex subgroups of employees.* Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1969.

References

Boehm, V. R. Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology*, 1972, *56*, 33–39.

Boehm, V. R. Differential prediction: A methodological artifact? *Journal of Applied Psychology*, 1977, *62*, 146–154.

Bray, D. W., & Moses, J. L. Personnel selection. *Annual Review of Psychology*, 1972, *23*, 545–576.

Cronbach, L. J. Equity in selection—Where psychometrics and political philosophy meet. *Journal of Educational Measurement*, 1976, *13*, 31–42.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. Ethnic group membership as a moderator of the prediction of job performance. *Personnel Psychology*, 1971, *24*, 609–636.

Fox, H., & Lefkowitz, J. Differential validity: Ethnic group as a moderator in predicting job performance. *Personnel Psychology*, 1974, *27*, 209–223.

Gael, S., & Grant, D. L. Employment test validation for minority and nonminority telephone company service representatives. *Journal of Applied Psychology*, 1972, *56*, 135–139.

Gael, S., Grant, D. L., & Ritchie, R. J. Employment test validation for minority and nonminority clerks with work sample criteria. *Journal of Applied Psychology*, 1975, *60*, 420–426. (a)

Gael, S., Grant, D. L., & Ritchie, R. J. Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology*, 1975, *60*, 411–419. (b)

Grant, D. L., & Bray, D. W. Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology*, 1970, *54*, 7–14.

Gross, A. L., & Su, Wen-huey. Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology*, 1975, *60*, 345–351.

Hunter, J. E., & Schmidt, F. L. A critical analysis of the statistical and ethnical implications of five definitions of test fairness. *Psychological Bulletin*, 1976, *83*, 1053–1071.

Hunter, J. E., & Schmidt, F. L. Differential and single-group validity of employment tests by race: A critical analysis of three recent studies. *Journal of Applied Psychology*, 1978, *63*, 1–11.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. *Journal of Applied Psychology*, 1977, *62*, 245–260.

Katzell, R. A., & Dyer, F. J. Differential validity revived. *Journal of Applied Psychology*, 1977, *62*, 137–145.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. *Testing and fair employment.* New York: New York University Press, 1968.

Lefkowitz, J. Differential validity: Ethnic group as a moderator in predicting tenure. *Personnel Psychology*, 1972, *25*, 223–240.

Lopez, F. M. The industrial psychologist: Selection and equal employment opportunity (a symposium): III. Current problems in test performance of job applicants: I. *Personnel Psychology*, 1966, *19*, 10–18.

Mitchell, M. D., Albright, L. E., & McMurry, F. D. Biracial validation of selection procedures in a large southern plant. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, *3*, 575–576. (Summary)

Novick, M. R., & Petersen, N. S. Towards equalizing educational and employment opportunity. *Journal of Educational Measurement*, 1976, *13*, 77–88.

O'Connor, E. J., Wexley, K. N., & Alexander, R. A. Single-group validity: Fact or fallacy? *Journal of Applied Psychology*, 1975, *60*, 352–355.

Ruda, E., & Albright, L. E. Racial differences on selection instruments related to subsequent job performance. *Personnel Psychology*, 1968, *21*, 31–41.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology*, 1973, *58*, 5–9.

Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist*, 1974, *29*, 1–8.

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 1977, *62*, 529–540.

Schmidt, F. L., & Hunter, J. E. Moderator research and the law of small numbers. *Personnel Psychology*, 1978, *31*, 215–232.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, 1976, *61*, 473–485.

Shuey, A. M. *The testing of Negro intelligence.* Lynchburg, Va.: J. P. Bell, 1958.

Toole, D. L., Gavin, J. F., Murdy, L. B., & Sells, S. B. The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology*, 1972, *25*, 661–672.

Wollowick, H. B., Greenwood, J. M., & McNamara, W. J. Psychological testing with a minority group population. *Proceedings of the 77th Annual Convention of the American Psychological Association*, 1969, *4*, 609–610. (Summary)

Received February 17, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 736–765

Unfair Discrimination in the Employment Interview: Legal and Psychological Aspects

Richard D. Arvey
University of Houston

The psychological and legal literature concerning evidence of bias or unfairness in the employment interview with regard to blacks, females, handicapped persons, and the elderly is reviewed. This review indicates that (a) the interview is highly vulnerable to legal attack and one can expect more future litigation in this area; (b) the mechanisms and processes that contribute to bias in the interview are not well specified by researchers; (c) findings based predominantly on resume research show that females tend to receive lower evaluations than males, but this varies as a function of job and other situational characteristics; (d) little evidence exists to confirm the notion that blacks are evaluated unfairly in interview contexts; (e) a relative dearth of research exists investigating interview bias against the elderly and handicapped individuals; and (f) evidence concerning the differential validity of the interview for these minority and nonminority groups is virtually nonexistent. A number of research needs and directions are specified.

Despite research that indicates that the employment interview has limited reliability and validity (Mayfield, 1964; Ulrich & Trumbo, 1965; Wright, 1969), organizational use of the interview in helping to make selection and promotion decisions persists. The statement of Dunnette and Bass (1963) that the personnel interview is the most widely used method of selecting employees still holds true today. In fact, there is some speculation that the employment interview may be gaining in popularity because of increased court and legal pressures brought to bear on employers' pencil-and-paper testing practices. In view of the increased likelihood of their employment tests' being subjected to legal scrutiny, employers are dropping the use of tests and placing even more reliance on the interview as the major decision-making tool.

An important but unsettling question remains, however, for those who abandon tests and adopt the interview for use in selecting and promoting employees. Does use of the interview result in unfair discrimination against protected minority group members? This question becomes even more important when one realizes that the interview is considered a *test* under the Equal Employment Opportunity Commission (1970) testing guidelines for selecting employees. Thus, the employment interview is (and has been) subjected to judicial review. Moreover, some recent evidence indicates that the interview process tends to yield judgments and evaluations about minority group members that make it less probable that they will be hired or promoted than that nonminority group members will be hired, even though the members of the two groups have substantially equal qualifications.

I intend to accomplish several objectives in the present article: First, I discuss various definitions of unfair discrimination in general and how components of these definitions apply to the employment interview; second,

---

This article is a greatly expanded version of a chapter in Arvey (1979).

I would like to thank Sheldon Zedeck for his review of earlier drafts of the article. A large portion of the research was completed while I was a visiting professor at University of California, Berkeley.

Requests for reprints should be sent to Richard D. Arvey, Department of Psychology, University of Houston, Central Campus, Houston, Texas 77004.

I review the legal literature and case law regarding unfair discrimination in the interview; third, I review the basic processes and mechanisms by which minority group members may come to receive lower evaluations based on the interview procedure; fourth, I review and summarize the research literature available concerning the evaluations given to minorities as a result of the interview process and whether these judgments exhibit differential validity in predicting job success; finally, I summarize trends and offer suggestions for future research efforts.

I selectively focus on the legal and research literature that concerns four basic minority groups: blacks, females, the elderly,[1] and the handicapped.[2] I have chosen to deal primarily with these particular groups because (a) the existing data base deals almost exclusively with these particular subgroups and (b) because the principles exemplified within the article generalize reasonably well to other subgroups.

Unfair Discrimination: Background and Definitions

As many of us are undoubtedly aware, Congress passed the Civil Rights Act in 1964, and Title VII of this act forbids discrimination against individuals in employment settings based on factors of race, national origin, religion, or sex. The agency charged with the interpretation and enforcement of this act was the Equal Employment Opportunity Commission (EEOC). In 1970 and again in 1978, this commission issued interpretative guidelines concerning the use and validation of employment tests in organization settings to insure the fairness of selection devices.[3] As mentioned, the employment interview was explicitly included under the rubric of tests and was subject to these same guidelines. Judicial definition of unfair discrimination gained greater precision in the now famous *Griggs* v. *Duke Power Company*[4] Supreme Court decision. In this decision, the Supreme Court ruled that the use of a general intelligence test and the use of a mechanical aptitude test were illegal because their use resulted in more whites than blacks being hired (adverse impact) and the test was not shown to be valid. The Court clearly enunciated a *shifting-burden-of-proof* principle, which applies to almost all litigation in this area today. First, an individual filing suit must demonstrate an adverse effect due to the use of a particular selection device.[5] For example, the finding that significantly more minority members than nonminority members were rejected for a position on the basis of an interview might be sufficient evidence to establish adverse impact. Subsequently, the burden of proof shifts to the employer, who must then establish the validity of the selection device. Since the interview is considered a form of test, the same two-pronged procedure is applicable. For an employment interview to be judged unfair, first, evidence would need to be presented that minority group members were evaluated more poorly than nonminority members on the basis of the interview and that these evaluations resulted in an adverse impact, and second, the employer would have to fail in establishing the validity of the interview.

It should be noted that another channel by which minority group members might challenge employment selection practices and the interview is by filing suits under the equal protection clauses of the Fifth and Fourteenth Amendments of the U.S. Constitution. Prior to 1975, suits filed under these authori-

---

[1] Individuals within the ages of 40 and 70 were extended protective coverage in the Age Discrimination in Employment Act of 1967 and recent amendments.

[2] Handicapped individuals were extended coverage by the Vocational Rehabilitation Act of 1973.

[3] In the past, other agencies (e.g., the Office of Federal Contract Compliance, Justice Department, etc.) issued separate testing guidelines. More recently, the various agencies jointly issued uniform guidelines on employee selection procedures (Equal Employment Opportunity Commission et al., 1978).

[4] Griggs v. Duke Power Company, 3 FEP 175 (1971).

[5] It should be noted that the more recent uniform guidelines stress that evidence for adverse impact should be assessed on the basis of the *total* selection process (the "bottom line" approach) in contrast with reviewing a single component of the process (i.e., the interview). It is not yet clear which reviewal process will be accepted by the courts.

ties were reviewed under the same standards as Title VII suits. However, in 1975, the Supreme Court ruled (in *Washington* v. *Davis*[6]) that when a suit is filed under these amendments, plaintiffs must demonstrate that the employer had an intent to discriminate (which is typically quite difficult to prove).

The 1970 EEOC guidelines call for the examination of whether a testing device (i.e., interview) exhibits the same pattern of validity for both majority and minority group candidates. Thus, an employer must determine that the interview is equally valid for both segments of the applicant pool. Employers are asked to determine whether their employment devices are *differentially valid* (Boehm, 1972). However, the more recent uniform guidelines (Equal Employment Opportunity Commission et al., 1978) indicate that instead of simply having researchers determine whether differential validity exists, studies of "fairness" should be carried out when feasible. One more sophisticated psychometric model of fairness calls for the investigation of regression line differences between majority and minority candidates (Cleary, 1968); that is, for the interview to be fair to minority and nonminority group members, the regression lines computed should be equal.

The more complex psychometric models of test fairness need not be presented fully here for the purposes of this article for two reasons: First, little agreement exists today concerning which of the more complex models of fairness is the most adequate model. Several relatively complex models have been advanced (Cole, 1973; Petersen & Novick, 1976; Thorndike, 1971) and are currently being evaluated and compared (Hunter & Schmidt, 1976). Second, research on interview fairness has simply failed to examine the issue from the perspective of these more sophisticated psychometric models, although it would be appropriate to do so.

To summarize, the fairness of the interview from a legal perspective can be evaluated from basically two perspectives: from the degree of adverse impact shown by interview judgments and from the degree of validity or job relatedness of the interview for majority and minority candidates.

Legal Aspects of the Interview

How has the employment interview fared when challenged in court by individuals claiming that they were discriminated against unfairly as a result of an interview? In almost all cases, the shifting-burden-of-proof standard is used. Cases in which the use of the interview in employment decisions has been challenged are shown in Table 1.

Perhaps the first and most often cited case in this area is *Rowe* v. *General Motors*,[7] decided by the Fifth Circuit Court of Appeals. The decision actually had to do with the performance appraisal used by the company in determining promotions, but because the case is used frequently as a controlling case, it is important to review it here. The decision dealt with the subjective nature of employment decisions, an obvious component of the interview. In this situation, foremen's subjective evaluations of hourly employees' ability, merit, and capacity were used in making promotion decisions. After determining that adverse impact had occurred regarding blacks, the court felt that the performance appraisal system violated Title VII in several ways: First, foremen were given no written instructions pertaining to the qualifications necessary for promotion; second, the standards that were determined to be in control were vague and subjective. In summarizing the decision, the court added that

> all we do today is recognize that promotion/transfer decisions which depend almost entirely upon the subjective evaluation and favorable recommendation of the immediate foreman are a ready mechanism for the discrimination against blacks, much of which can be covertly concealed. (*Rowe* v. *General Motors*, p. 450)

Thus, this decision casts some doubt on the employer's use of a subjective decision-making process if such a process resulted in adverse impact.

---

[6] Washington v. Davis, 12 FEP 1415 (1976).
[7] Rowe v. General Motors, 4 FEP 445 (1972).

Table 1
Litigation Concerning Interviews

Case	Alleged discrimination against	Adverse impact shown	Validity of interview shown	Comment
Rowe v. General Motors (1972)	Blacks	Yes	No	Subjective standards said to be ready mechanism for discrimination
Equal Employment Opportunity Commission Decision No. 72-0703 (1971)	Blacks	Yes	No	Hiring system must permit review
Hester v. Southern Railway Company (1974)	Blacks and females	No		Interview noted to be subjective
United States v. Hazelwood School District (1976)[a]	Females	Yes	No	
Weiner v. County of Oakland (1976)	Females	Yes	No	
Harless v. Duck (1977)	Females	Yes	Yes	Interview successfully defended on the basis of content validity
King v. New Hampshire Department of Resources and Economic Development (1977)	Females	Yes	No	Questions in interview were indicative of discriminatory intent
Bannerman v. Department of Youth (1977)	Females	No		Adverse impact not shown

[a] Later appealed to the Supreme Court, but issues associated with the interview were not a part of this later decision.

In a 1971 decision,[8] the EEOC ruled that an employer's decision not to hire a black woman because of her "poor attitude" during the interview was in violation of Title VII. In this case, the EEOC cited a number of court decisions to establish that if discriminatory impact of a hiring system is shown, "it is essential that the system be objective in nature and be such as to permit review" (EEOC Decision No. 72-0703, p. 437).

The employment interview was also given court scrutiny in *Hester* v. *Southern Railway Company*.[9] A black female applicant was denied a clerical job partly as a result of the interview process. Although the district court ruled that an adverse impact had been demonstrated and that the interviewing procedure was faulty because of its subjective nature and because it was based on "no formal guidelines, standards and instructions," the court of appeals overturned the decision because there was no clear proof that these selection procedures (tests, interviews, etc.) had resulted in adverse impact.

A more recent case involved a court decision that struck down the use of the interview used in hiring teachers. In *United States* v. *Hazelwood School District*,[10] the court noted that the subjective interview process used in making selection decisions was similar to the vague and subjective criteria used in *Rowe* v. *General Motors*. It is instructive to read what the court said in this instance:

> Principals are free to give whatever weight they desire to subjective factors in making their hiring decisions. Indeed, one principal testified that interviewing an applicant was "like dating a girl, some of them impress you, some of them don't." . . . No evidence was presented which would indicate that any two principals apply the same criteria—objective and subjective—to evaluate applicants. (*United States* v. *Hazelwood School District*, p. 7576)

An interesting 1976 case (*Weiner* v. *County of Oakland*[11] dealt specifically with the kinds of questions asked in the interview and their possible bias. Mrs. Weiner applied for the position of intermediate planner and was given an oral interview that apparently was scored in some systematic way. Although Mrs. Weiner was ranked third on the list of eligible applicants, four men were hired to fill the available positions. As grounds for not hiring Mrs. Weiner, the county was only able to suggest some doubt about the flexibility of her approach to planning.

The court ruled that Mr. Weiner had demonstrated an adverse impact. At this point, the burden of proof shifted to the county, which had to prove that it had valid business requirements justifying its conduct. The county attempted to defend the use of the interview by asserting that the decision reached was based on subjective evaluations made during the interview that were in no way the product of sex discrimination.

The court, however, reviewed the kinds of questions that were asked of Mrs. Weiner and found that they were suggestive of bias against women. Questions such as whether her husband approved of her working, whether her family would suffer if she were not home to prepare dinner, and whether she was able to work compatibly with young, aggressive men were asked. The court ruled that these kinds of questions during the interview, along with other facts, were sufficient to substantiate the charge of discrimination and awarded back pay and attorney's fees.

A recent case provided sufficient evidence to the court regarding the interview to survive challenge. *Harless* v. *Duck*[12] involved a situation in which a woman brought a class action suit against a midwestern police department charged with discrimination in hiring because of sex. The department used, along with several other tests, a structured oral interview that consisted of approximately 30 questions posed to each candidate by a team of interviewers. The questions were designed to determine an "applicant's communication skills, decision-making and problem-solving skills, and reactions to stress

[8] Equal Employment Opportunity Commission Decision No. 72-0703, 4 FEP 435 (1971).
[9] Hester v. Southern Railway Company, 8 FEP 646 (1974).
[10] United States v. Hazelwood School District, 11 EPD 10854 (1976).
[11] Weiner v. County of Oakland, 14 FEP 380 (1976).
[12] Harless v. Duck, 14 FEP 1616 (1977).

situations." It was determined that 43% of the females failed the oral interview, compared with 15% of the males. After some discussion of proper sample sizes for detecting significant differences, the court determined that the discrepancy in pass rates was significant and that the interview did indeed have a discriminatory or adverse effect. In defending the validity of the interview, the organization relied on two sources of evidence:

1. The oral interview had construct and content validity. The expert witness for the organization testified that the structured oral interview portions of the exam, which consisted of hypothetical questions simulating situations likely to be encountered by patrolmen, measured several dimensions identified through job analysis that differentiate among persons who would be better patrol officers if put in a position to perform patrol functions.

2. A significant relationship between performance in the interview and performance at the police academy was shown. A previous Supreme Court decision (*Washington* v. *Davis;* see Footnote 6) had affirmed the use of measures of training success as legitimate criteria against which to validate a selection instrument. The court found this evidence sufficient to demonstrate the validity of the interview.

Another recent case is *King* v. *New Hampshire Department of Resources and Economic Development,*[13] in which a court of appeals found that the questions asked of a female who applied for the job of a state meter patrol officer helped to establish that discriminatory intent had occurred. In this instance, a female applicant was asked "whether she could wield a sledge hammer, whether she had any construction industry experience, and whether she could 'run somebody in'" (p. 670), none of which were related to the job in question. The court indicated that the employer's discriminatory intent was proved largely by its own words and actions.

Finally, in *Bannerman* v. *Department of Youth Authority*[14] the use of a panel interview was challenged. Candidates applying for a parole agent job were interviewed by a panel of three interviewers who were asked to judge each candidate in relation to stated "critical class requirements" (e.g., demonstrated ability to relate to youths and to gain their respect and confidence). However, the plaintiffs were not able to demonstrate to the satisfaction of the court any discriminatory bias against women; that is, there was no statistically significant difference between the pass rates of the males and females interviewed by the panel.

It is somewhat surprising that more cases dealing with interviews have not been litigated. It seems apparent that one direction in which the courts are moving is toward the exploration of the nature of and kinds of questions asked and the information elicited in the interview in more depth; that is, the content of the interview is being more fully examined. For example, inquiries during the interview that might convey to the applicant the impression that persons in a protected class will be discriminated against will now be viewed as discriminatory. In one case (cited in Babcock, Freedman, Norton, & Ross, 1975), the EEOC and the New York Human Rights Commission concluded that a New York law firm had violated Title VII when interviewers emphasized to female applicants that the firm had only one female lawyer and that she was assigned to an area of work traditionally performed by women. The conclusion was that "the interviews are conducted in such a manner as to express a preference for men and to discourage women from pursuing employment with respondent firm" (Babcock et al., 1975, p. 380).

Among these same lines, the specific kinds of information elicited on application forms and during interviews are being litigated. Managers are currently confused about what they may or may not ask during an interview. Although I do not provide a review of these more specific interview inquiries, it should be noted that litigation revolves around two basic themes:

---

[13] King v. New Hampshire Department of Resources and Economic Development, 15 FEP 669 (1977).

[14] Bannerman v. Department of Youth Authority, 17 FEP 820 (1977).

1. Do particular kinds of questions convey an impression of an underlying discriminatory attitude or intent? That is, references to "girls" and inquiries into non-job-related areas such as marital status, parenthood, child care, and so on, when these same questions are not presented to male candidates, may be sufficient to convince a court that discriminatory animus or intent is operating.

2. Does the inquiry operate in a way that demonstrates a differential impact or adverse effect on protected groups? If so, is the particular information valid or job related? Thus, organizations should avoid interview questions that operate in such a way as to differentially affect minority groups, unless such questions are job related.

Guidelines concerning preemployment inquiries have been set forth by a variety of state human rights commissions as well as by a set of guidelines issued by the EEOC. For example, the Washington State Human Rights Commission (1979, pp. 2923–2926) stated the following to be unfair and illegal preemployment inquiries when they cannot be shown to be job related: (a) all inquiries related to arrests; (b) any inquiry concerning citizenship; (c) specific inquiries concerning spouse, spouse's employment or salary, children, child care arrangements, or dependents; (d) overgeneral inquiries (e.g., "Do you have any handicaps?"—which would tend to divulge handicaps or health conditions that do not relate to fitness to perform the job); (e) whether the applicant is married, single, divorced, engaged, widowed, or any other inquiry as to marital status; (f) type or condition of military discharge; (g) any questions related to pregnancy; and (h) whether applicant owns or rents a home.

To summarize, although there have not been an overwhelming number of lawsuits involving the discriminatory nature of the employment interview, the litigation that has evolved clearly indicates that the interview is vulnerable to such suits. Interviews will indeed be treated like tests and reviewed by these same standards. I predict a greater number of suits in this area during future years. Organizations may find themselves even more ill equipped to defend the interview because of the little attention paid to quantifying interview judgments or conducting research to determine the reliability, validity, or adverse effects of the interview process.

Processes Inherent in Interview That Potentially Contribute to Differential Evaluation

The interview process is highly subjective. Schmidt (1976) has summarized a myriad of factors that contribute to interviewers' judgments and evaluations. Because of its basic subjective nature, the interview process is vulnerable to the personal biases, prejudices, and stereotypes of interviewers, thus making it open to challenge from civil rights litigants. Interviewers may form poorer evaluations of minority group members than of nonminority candidates, even when the candidates are substantially similar with regard to their job qualifications. Just how or why these differential evaluations are made is not well-known. There do appear to be two possible mechanisms that contribute to differential evaluations: (a) stereotyping and (b) differential behavior emitted during the interview.

Stereotyping

Decisions that are made on the basis of the interview are generally subjective in nature and thus susceptible, according to some individuals, to the influence of stereotypes. However, Brigham (1971) indicated that there is a good deal of confusion concerning the precise definition of stereotypes and their correlates. Most definitions (e.g., Lippman, 1922) revolve around the notion that stereotyping involves making judgments about people on the basis of their membership in a particular group. Once an individual's membership in a particular class or category is established (e.g., race, sex, age, etc.), a number of trait characteristics are ascribed to the individual based on the traits associated with the larger class of which he or she is a member. Thus, stereotyping involves basically two processes: (a) the for-

mation of impressions and trait descriptions of particular classes and categories of individuals and (b) the assignment of these traits to a particular individual once his or her membership in that class or category is known.

For example, Rosen and Jerdee (1976a) indicated that common stereotypes ascribed to males include such trait descriptions as adventurous, competitive, objective, dominant, decisive, and rough, whereas females are commonly described as compassionate, dependent, submissive, emotional, and so forth.

While many researchers have characterized stereotypes as the product of a rigid and faulty reasoning process that operates to help individuals rationalize and justify their hostility and prejudices, Brigham (1971) and Hamilton (1976) have suggested that stereotypes are not unlike other normal cognitive functioning processes. Hamilton indicated that stereotypes are "unfounded overgeneralizations," in that the perceiver does not use the available information in an optimal manner and bases his or her conclusions about a particular social group on poor evidence. In a similar vein, Katz (1960) has delineated several useful functions that stereotypes fulfill for different people (instrumental, ego-defense, value-expressive, and order functions).

It is interesting to note that although the notion of stereotyping is frequently invoked to explain why differential evaluations occur during interviews, the precise nature of how stereotypes operate and produce these different evalutions is not well specified. There appear to be at least three lines of speculation concerning this process. First, the stereotypes or trait characterizations may be essentially negative in nature; that is, the traits ascribed to a minority group and individuals from that group may consist of basically negative tones (e.g., blacks are dirty, uneducated, unintelligent, etc.). Thus, essentially negative attitudes and opinions concerning particular minority groups may have their basis in these stereotypic characterizations. Individuals holding essentially negative opinions and attitudes toward minority group members might be expected to give lower evaluations in an interview because of these attitudes. For example, Terborg and Ilgen (1975) found that attitudes toward women correlated significantly with a subject's decision to hire a female engineer. Individuals with essentially negative attitudes about women were less likely to give favorable evaluations than were those with more positive attitudes.

Similarly, Britton and Thomas (1973) gathered opinion data from 56 employment interviewers and found that they ranked older individuals as more difficult to place and train and as more slow in maintaining production. In addition, females were seen as more likely to be absent from work and as less skilled than men. Finally, older women were viewed as having fewer saleable skills than older men, but younger men and women (aged 18 years) were seen as having basically the same saleable skills. These negative opinions may lead to lower evaluations of older individuals and women in interview situations.

A second and more indirect manner by which stereotypes may affect interview evaluations is based on the process of matching the stereotypic traits with the characteristics thought to be necessary to perform the job. If a relatively poor match results, an interviewer may reject the candidate outright or give lower evaluations. Note that this process does not necessarily imply that the stereotypic traits are negative—indeed, they may even be positive in nature. Evaluations are potentially inaccurate (a) because of the inaccuracy of the stereotypes attributed to the individual, (b) because of the inaccuracy of the characteristics deemed necessary to perform the job, or (c) because the matching process itself is a poor decision strategy.

The research conducted by Schein (1973, 1975) is particularly representative of this second viewpoint. Schein (1973) asked 300 male managers to indicate which of 92 adjectives best described (a) women in general, (b) men in general, or (c) successful middle managers (each manager described only one of these three subgroups). She found that the relationship between the average descriptions of the middle manager

and the average description of the males was much higher ($r = .62$) than the relationship between the descriptions of the managers and of the females ($r = .06$). In fact, on 60 of the 92 items, the descriptions of the managers were more similar to the descriptions given to men than to the descriptions given to women. These findings suggest that even before an interview or any formal selection process has begun, the perceived similarity between the characteristics of successful managers and of men in general increases the probability of a male's rather than a female's being given a higher evaluation. These results were also replicated with a group of 167 female managers who also described the various subgroups (Schein, 1975). In this study, 167 female middle managers rated women in general, men in general, or successful middle managers on the same 92 adjectives. Like the male managers in the previous study, the female managers provided descriptions of successful middle managers that were far more similar to descriptions of men than to descriptions of women. The results suggest that female managers are as likely as male managers to accept stereotypical male characteristics as the basis for success in management.

A third manner by which stereotypes possibly operate is to shape the kinds of expectations and standards that interviewers have of job candidates during the interview. For example, it may be that an interviewer, after learning that the person to be interviewed next is female, evaluates the candidate on a different set of criteria—for example, beauty, typing skills, poise, and so on—than used for evaluating a male candidate. A study by Cecil, Paul, and Olins (1973) attempted to identify the qualities perceived to be important for male and female applicants for the same job. Over 100 subjects indicated the importance of each of 50 variables (e.g., pleasant voice, expressive self-will, etc.) to interviewers evaluating either a male or a female job candidate.

The results indicated that the kinds of standards and criteria used to evaluate the candidates depend on whether an applicant is male or female. The standards used to evaluate females are more clerical and cosmetic in nature, whereas the standards for males are along aggressive and persuasive dimensions.

Stereotypes obviously are also associated with race. Despite the Civil Rights Act of 1964, stereotyping of blacks persists today. For example, Karlins, Coffman, and Walters (1969) compared the racial stereotypes of undergraduate students in 1967 with the stereotypes reported in 1933 and 1951. Subjects in all three studies indicated those traits thought to be typical of blacks. Although the data indicated a consistent trend over the 25-year period toward giving more favorable characterizations of blacks, a considerable amount of stereotyping still exists. The students in the study described blacks as lazy, superstitious, musical, ostentatious, and pleasure loving. Considerably more information concerning the nature of racial and ethnic stereotypes can be found in Jones (1972) and Brigham (1971).

The nature of age stereotypes and how they relate to the job have recently received some research attention. Rosen and Jerdee (1976c) asked over 100 business students and realtors to imagine that they were going to meet two individuals for the first time and that the only information they had about the two people was that one was 30 years old and the other 60 years old. The participants in the study then considered 65 characteristics and indicated the degree to which each characteristic described the average 60-year-old male and the average 30-year-old male. Significant differences were found on several characteristics and are summarized as follows: The 30-year-old male was described as significantly more productive, efficient, and motivated, more capable of working under pressure, more ambitious, eager, and future oriented, more receptive to new ideas, more adaptable, and more versatile. The 60-year-old male was described as significantly more accident prone and more rigid and dogmatic.

Clearly the traits and characteristics ascribed to the older group are not particularly favorable. Rosen and Jerdee (1976c) noted that these stereotypes are also not consistent with objective research findings and

changes associated with aging (Schaie, 1974).

To summarize, the specific nature of stereotypes that interviewers hold concerning different minority groups may well influence their evaluations of these candidates during the interview process. To the extent that the stereotypes are basically negative, deviate from the perception of what is needed for the job, or translate into different expectations and standards of evaluation for minority group members, stereotypes may well have the effect of lowering the evaluations of interviewers, even when the candidates are equally qualified for the job.

Terborg (1977) has sounded a warning against the too quick acceptance of the notion that the pervasive effects of stereotypes explain or account for all differences in treatment by men and women managers. He suggested alternative explanations that may account for apparent discrimination. One alternative he posed is that many women are truly not yet qualified for managerial positions: "This is not meant to suggest that women lack potential for the job, but that the cumulative effects of past discrimination have prevented women from gaining the necessary skills and experience" (Terborg, 1977, p. 649).

Differential Behavior of Interviewers

One additional explanation for different evaluations of minority group candidates as a result of interviews is that minority applicants may behave in a manner that seems different and unfamiliar to interviewers. Hall (1966) specifically argued that subcultural differences in nonverbal behavior have resulted in whites' misreading of black applicants and therefore in blacks' failure to get jobs; that is, it is possible that blacks emit verbal and nonverbal behaviors (e.g., jive talk) that are acceptable and even desirable in their subculture, but that these same behaviors are misinterpreted or confused by a white interviewer. An interesting study was conducted in this regard by Fugita, Wexley, and Hillery (1974). White and black (20 of each) female undergraduate students were asked to participate in an employment interview. Acting as the interviewers were 2 black and 2 white males; eight questions were asked of the interviewees. Results of the study showed that the black interviewees maintained significantly less eye contact with both white and black interviewers. The least amount of visual interaction occurred when a black interviewee was interviewed by a black. Black interviewers were also given shorter glances than were white interviewers by both black and white interviewees. In sum, the race of the interviewee and of the interviewer appeared to be significant factors in the kind and amount of behavior that occurred during the interview. What would have also been of great interest is whether these differences in behavior also produced differences in evaluations by the interviewers.

It is possible that women, older workers, and handicapped individuals also differ in their reactions during interviews, which could contribute to lower evaluations. For example, older job candidates may be more thoughtful and cautious in their interview responses than are younger candidates. Sterrett (1978) conducted a study in which 100 male and 60 female managers evaluated videotapes of a male applicant in which the intensity of nonverbal body cues was manipulated. Results indicated that the male and female interviewers reacted significantly differently to the different kinds of body language.

Research Findings Concerning Different Evaluations in the Employment Interview

In the previous section, it was learned that one component of unfair discrimination from a legal perspective is whether the evaluations stemming from the interview result in an adverse impact on protected groups. In the present section, the various research studies that have investigated whether equally qualified minority group members receive lower evaluations on the basis of interviews are reviewed. These studies have generally employed one of three kinds of research strategies.

Resume Studies

In this kind of study, subjects (students, managers, recruiters, etc.) are asked to review a series of job resumes and to determine the suitability of each of the candidates for employment and/or the starting wage that might be offered. Each resume usually contains information about educational background and past experience, with a glossy photograph attached. In the typical study, the race, sex, age, or handicapped status of the job candidates can be manipulated through the photograph and the name, which is also printed on each resume. Thus, half the interviewers might be asked to make evaluations concerning, for example, five males and five females. The other half of the interviewers might be asked to evaluate five males and five females, with the only change being that the names and photographs are switched so that no change whatsoever occurs with regard to the qualifications of the candidates; the only changes made are in regard to sex, race, and so on. Obviously, the interviewers are unaware that the resumes they are evaluating may differ from those being evaluated by other interviewers. In addition, some studies have added several other variables such as applicant attractiveness, type of job, and so forth to determine if these characteristics may interact somehow with minority status and thus influence the evaluations given.

Another strategy researchers use is to have subjects evaluate and rate several resumes that vary by sex, race, and so on. Care is taken to equate the resumes with relevant characteristics other than the variables studied. One potential problem, then, with this within-subject design is the degree to which potentially confounding variables are effectively controlled.

Because many of the studies in this area use this resume approach, they obviously do not involve face-to-face interviews with real people. Instead, subjects are asked to evaluate "pencil-and-paper" people for jobs in organizations that are often not well described. One must infer that any discriminatory effects found in these studies generalize to actual interview circumstances.

"In-Basket" Study

In this design, participants in the study (students, managers, etc.) are asked to assume the role of a personnel director or manager who works through an "in-basket" and must take action and react to a number of items in memorandum or letter form. Typically, each in-basket provides information about the various departments in the organization and information about the members of the organization (e.g., performance appraisal data, attendance information, etc.). Also contained in each participant's in-basket are a series of different types of personnel problems. The participants are asked to make some kind of decision based on the information given. One of the decisions presented is the hiring or promotion of a particular individual. The problems are written in two or more versions so as to change the sex, race, age, and so on of one or more of the characters in the problem. For example, one participant may receive an in-basket set in which the problems involve whether a male should be hired or promoted, whereas another participant may receive the exact same information, except that the sex of the person being considered for hire or promotion is female.

Other kinds of problems are presented in addition to hiring and promotion. For example, training and development opportunities, leaves of absences, job assignments, and so forth are investigated with regard to whether participants make different kinds of decisions based on the sex, race, and so on of the individuals depicted in the in-basket materials.

Videotape Studies or Field Experiments

Less frequent studies employ designs in which actual minority and nonminority interviewees are observed by interviewers either face to face or in videotape presentations. Typically, interviewers (subjects, managers, etc.) interview or observe only one job candidate, who is either a minority or a nonminority member, and make evaluations about the suitability of the candidate for hire. Efforts are made to control the content

of the interview to insure that the same questions are asked and that similar responses are delivered by the interviewees.

Table 2 summarizes the results of 23 studies that have investigated bias in the interview. As can be seen, most studies employed the resume approach, the second most frequent strategy was the in-basket strategy, and finally, the videotape or in vivo type of study was used in only three instances. Although there is considerable variability in the sample sizes used across the various studies, the majority of designs (19) included more than 75 subjects. Thus, the studies were in general reasonably powerful with regard to detecting significance (Hayes, 1973).

Applicant Sex

The studies in Table 2 show in a reasonably consistent fashion that females are generally given lower evaluations than males where these candidates have similar or identical qualifications.

Dipboye, Fromkin, and Wiback (1975) gave 30 male professional interviewers and 30 male undergraduates 12 resumes to evaluate and rate for the position of head of a furniture department. Applicants' sex, physical attractiveness, and scholastic performance were manipulated. Among other findings, a main effect ($F = 15.84$, $p < .01$) for applicant sex was found in which male applicants were evaluated higher than were female applicants. However, this effect accounted for only 1% of the rating variance. One potential confound in this study was that the description of the furniture department job included several references to "him," which perhaps connoted to subjects that male applicants were preferred.

Dipboye, Arvey, and Terpstra (1977) provided a partial replication and extension of the aforementioned study. College students ($N = 110$) evaluated 12 resumes for a management-trainee sales position. Besides the qualifications of the ratees, subjects' sex and attractiveness were investigated. In addition to other findings (summarized below), applicant sex demonstrated a significant main effect on an employability rating. Raters expressed significantly more willingness to hire a male than to hire a female applicant ($F = 20.44$, $p < .001$). Again, however, this effect accounted for a minute proportion of the variance in ratings (.006).

Haefner (1977) asked 286 managers to rate 16 resume profiles in which applicant characteristics varied by sex, age, race, and competence. All applicants were described as being disadvantaged. In addition to several interactions (discussed below), sex exhibited a significant main effect ($F = 78.53$, $p < .01$) in which males were given higher recommendations for hire than were females. A factor that might have influenced the results of the study is that the presentation of the stimulus profiles and subsequent evaluations took place over the telephone. The effects, if any, of using this data collection procedure are not known.

Rosen and Jerdee (1974a) used 235 male undergraduate subjects to evaluate male or female candidates for jobs with demanding versus routine job requirements. Females were rated significantly lower than were males on an overall rating ($F = 14.02$, $p < .01$).

Zikmund, Hitt, and Pickens (1978) varied sex and scholastic performance of applicants in a resume study in which 100 personnel directors were asked to reply to letters from the candidates that expressed interest in an accounting job. The dependent variable was the number of replies received. Females received a significantly lower number of replies ($\chi^2 = 4.5$, $p < .05$). In addition, the replies that were received by the females were significantly less positive in nature than were those for other candidates.

Several studies, however, failed to demonstrate main effects for applicant sex. Fidell (1970) asked 147 psychology department chairpersons to evaluate the chances of 10 candidates' getting an offer for a full-time position. Sex of the candidates was varied. Although the mean ratings of male and female candidates did not differ, women were offered positions at significantly lower levels than the positions offered males.

Using an in-basket methodology, Terborg and Ilgen (1975) asked 36 male under-

(*text continued on page 754*)

Table 2
Summary of Studies Investigating Discriminatory Effects in Interviews

Study	Minority variables investigated	Other variables investigated	Major criterion used	Interviewer status and sample size	Type of job	Results	Comment
				Resume Studies			
Fidell (1970)	Sex		Desirability of candidates	147 chairpersons of psychology departments	Academic position	Males = females	No difference in mean ratings, but significant effect found in which men offered higher ranking positions
Shaw (1972)	Sex and handicap	Type of job	Hire rating	64 college recruiters	Management trainee or engineer	Males > females, $t = 1.81$, $p < .05$ Handicapped = non-handicapped, $t = 1.78$, *ns*	Greater bias against females for management trainee job
Wexley & Nemeroff (1974)	Race	Perceived similarity of applicants to interviewer	Hire recommendation	120 college students	Mechanical-engineer technician	Whites = blacks, $F = .34$, *ns*	Perceived similarity of biographical background was significant determinant of interviewer evaluation
Rosen & Jerdee (1974a)	Sex	Job demands	Overall hire rating	235 college students	Demanding job or clerical job	Males > females, $F = 14.02$, $p < .01$ Sex × Job demands, $F = 6.27$, $p < .05$	Females evaluated more severely when job requirements were demanding and challenging
Dipboye, Fromkin, & Wiback (1975)	Sex	Qualifications and attractiveness of candidates	Hire rating	30 college students and 30 professional interviewers	Head of furniture store	Males > females, $F = 15.84$, $p < .05$, $\eta < .01$ Qualified > unqualified, $F = 185.40$, $p < .01$, $\eta^2 = .34$	Qualified and attractive candidates more likely to be hired, regardless of sex

Table 2 (*continued*)

Study	Minority variables investigated	Other variables investigated	Major criterion used	Interviewer status and sample size	Type of job	Results	Comment
Cohen & Bunker (1975)	Sex	Type of job	Hire recommendation	150 job recruiters	Editorial assistant or personnel assistant	Job × Sex, $\chi^2(1) = 16.416, p < .01$	Females more acceptable for editorial jobs than personnel jobs; opposite true for males
Krefting & Brief (1977)	Handicap	Experience	Suitability for hire	145 college students	Typist position	Disability × Experience, $F = 4.58, p < .05$	Disabled, inexperienced applicant evaluated higher than experienced, disabled applicant
Rose & Brief (Note 1)	Handicap	Type of job	Satisfy customers	145 college students	Public contact job or supervisory job	Job × Health Status, $F = 11.18, p < .01$	Epileptics expected to satisfy clients more if hired for nonpublic job versus public job
Haefner (1977)	Race, sex, and age	Competence	Hire rating	286 managers	Semiskilled position	Males > females, $F = 78.53, p < .01$ Younger > older $F = 103.00, p < .01$ Competent > incompetent, $F = 544.92, p < .01$ Race × Sex, $F = 4.68, p < .05$ Race × Age, $F = 5.00, p < .05$ Sex × Competence $F = 10.54, p < .01$ Age × Competence, $F = 11.71, p < .01$	Among highly competent applicants, males and younger candidates preferred

(*table continued*)

Table 2 (*continued*)

Study	Minority variables investigated	Other variables investigated	Major criterion used	Interviewer status and sample size	Type of job	Results	Comment
Dipboye, Arvey, & Terpstra (1977)	Sex	Qualifications, attractiveness of applicants, sex, and attractiveness of interviewer	Hire rating	110 college students	Sales trainee	Males > females, $F = 20.44$, $p < .01$, $\eta^2 = .006$ Qualified > unqualified, $F = 563.93$, $p < .01$, $\eta^2 = .50$	Raters given choice of one candidate chose highly qualified male significantly more often than highly qualified female
Cash, Gillen, & Burns (1977)	Sex	Type of job and attractiveness of applicant	Hiring potential recommendation	72 personnel directors	Masculine, feminine, or neutral	Males > females for masculine jobs, $F = 7.28$, $p < .01$ Females > males for feminine jobs, $F = 21.83$, $p < .01$	
Muchinsky & Harris (1977)	Sex	Rater sex and qualifications	Recommendation to hire	100 college students	Mechanical engineer, day care, and copy editor	Females > males for rater sex, $F = 6.77$, $p < .05$ Females > males, $F = 11.36$, $p < .05$ Qualified > unqualified, $F = 256.39$, $p < .05$	Low- or medium-qualified females given higher evaluations than males on two of the three jobs
Heneman (1977)	Sex	Qualifications (through test scores)	Job suitability	144 college students	Life insurance agent	Qualified > unqualified, $\omega^2 = .22$ Qualifications × Sex, $\omega^2 = .02$	Highly qualified females rated less suitable than highly qualified males

Table 2 *(continued)*

Study	Minority variables investigated	Other variables investigated	Major criterion used	Interviewer status and sample size	Type of job	Results	Comment
Rose & Andiappan (1978)	Sex	Sex of raters and sex of potential subordinates	Probability of success	75 college students	Management position	Female raters > male raters, $F = 3.91$, $p < .0$, $\omega^2 = .33$; Applicant Sex × Predominant Sex of Subordinate, $F = 13.22$, $p < .01$, $\omega^2 = .42$	Female raters evaluated applicants of both sexes more positively than male raters; females expected to be more successful when subordinates were predominantly female
Renwick & Tosi (1978)	Sex	Job type, marital status, undergraduate major, and graduate degree	Ratings of general suitability	64 male and 16 female graduate students	Management with extensive travel and management with home office duties	Sex × Undergraduate Major × Graduate Degree × Marital Status, $F = 4.53$, $p < .006$	No effects found for sex
Zikmund, Hitt, & Pickens (1978)	Sex	Scholastic performance	Number of replies to letter	100 personnel directors	Accounting	$\chi^2(1) = 6.0$, $p < .01$	Greater response rate to letters using initials than to letters from female applicants
Kryger & Shikiar (1978)	Sex	Sex of author of letter of recommendation and favorableness of recommendations	Would proceed with interview	75 personnel managers	Managerial job	Females > males, $F = 5.3$, $p < .05$	Interpreted in terms of reverse discrimination

(table continued)

Table 2 (*continued*)

Study	Minority variables investigated	Other variables investigated	Major criterion used	Interviewer status and sample size	Type of job	Results	Comment
				In-basket studies			
Rosen & Jerdee (1974b)	Sex	Job complexity	Recommendation to hire	95 bank supervisors	Complex or routine	Males > females, $\chi^2(1) = 6.53$, $p < .05$	Interaction between job complexity and sex not significant
Terborg & Ilgen (1975)	Sex		Decision to hire and starting salary	36 college students	Engineer	Males > females on starting salary, $F = 4.51$, $p < .05$	No difference between male and female applicants in decision to hire but females given lower salary
Rosen & Jerdee (1976b)	Age		Suitability for job	142 college students	Job requiring high-risk judgments	Younger > older, $t = 2.91$, $p < .01$	Older employees evaluated as more resistant to managerial influences, less promotable to position requiring creativity, less likely to approve older workers' requests for transfer to job requiring high physical activity, and less likely to recommend support of training programs for older employees

Table 2 (*continued*)

Study	Minority variables investigated	Other variables investigated	Major criterion used	Interviewer status and sample size	Type of job	Results	Comment
				Videotape or field studies			
Rand & Wexley (1975)	Race	Rater racial prejudice and need for affiliation	Hiring recommendation	160 college students	Mechanical-engineer technician	Biographical similarity, $F = 24.22$, $p < .01$	Race not significant; similarity between applicant and interviewer significantly related to evaluations; more similar candidates given higher evaluations
Johnson & Heal (1976)	Handicap		Job offer	50 counselors at employment agencies		Nonhandicapped > handicapped, $r = .278$, $p < .01$	Handicapped applicant received significantly fewer job offers after visiting employment agency
Dipboye & Wiley (1977)	Sex	Aggressiveness of candidate	Recommended for hiring	66 college recruiters	Supervisor in retail department store	Males = females, aggressive > passive, $F = 37.41$, $p < .01$	No difference between males and females in recommendation for hiring, but females significantly evaluated more favorably on likability, qualifications, and suitability for job

graduates to evaluate male and female job candidates and found that while there was no difference in hire rating on the basis of applicant sex, females were assigned a significantly lower starting salary than were identical male applicants ($F = 4.51$, $p < .05$). This study is somewhat limited, however, by the small sample size and the use of college undergraduates as subjects. Dipboye and Wiley (1977) showed videotapes of aggressive and passive male and female applicants to 66 college recruiters and found that while aggressive candidates were evaluated significantly higher than passive candidates on an employability scale, there were no differences between the male and female interviewees. However, the recruiters perceived the females' overall qualifications and their experience and training as superior to those of the males. One possible problem associated with this study is that only one particular stimulus male and stimulus female were presented in the videotape conditions. Thus, the observed effects may have been unique to the specific individuals presented.

Renwick and Tosi (1978) asked 80 male and female graduate students to evaluate resumes for two managerial jobs; all applicants were portrayed as exceptionally well qualified. The applicant variables manipulated were sex, marital status, undergraduate major, and graduate degree. Each subject reviewed and evaluated 10 profiles. No evidence of differential evaluations as a function of applicant sex was found, except as a component of a four-way interaction.

A study by Muchinsky and Harris (1977) yielded results somewhat opposite to the previous findings. College students ($N = 100$) evaluated resumes that varied by applicant sex and qualifications for three different jobs (mechanical engineer, day-care person, and copy editor). Although numerous two- and three-way interactions were observed, a main effect ($F = 6.77$, $p < .95$) for applicant sex occurred in which females were given higher ratings than males.

A similar finding was observed in a related study by Kryger and Shikiar (1978). Personnel managers ($N = 75$) evaluated letters of recommendation of male and female candidates for a management-trainee job. Sex of the author of the letter of recommendation was manipulated, as was the favorableness of the letter. Subjects rated applicants in terms of whether the applicant should be interviewed. Results indicated that subjects were far more likely to proceed with an interview with female applicants than with male applicants ($F = 5.3$, $p < .05$). Kryger and Shikiar interpreted their results as being a function of affirmative action considerations by personnel managers. It should be noted that this study has more to do with whether an applicant is granted an interview than with decisions resulting from the interview itself. Although more females may be granted interviews, the decisions resulting from actual interviews may be unfavorable for women, as noted above.

In addition, several studies have revealed that both male and female raters give lower evaluations to female applicants. Studies by Dipboye et al. (1975) and Dipboye et al. (1977) showed that the observed effects are not confined only to male interviewers. However, when sex of the rater was investigated in the Rose and Andiappan (1978) and Muchinsky and Harris (1977) studies, significant main effects were observed in which female raters gave significantly higher ratings to applicants of both sexes than did male raters. Thus, there is some indication that female raters are more lenient than male raters, which corresponds to the findings of London and Poplawski (1976), in which females were observed to give more lenient evaluations of stimulus objects in a study involving the formation of stereotypes. However, Renwick and Tosi (1978) found no evidence of a sex-of-interviewer effect.

In addition to simply searching for main effects, research has focused on several variables that are predicted to interact with applicant sex in influencing the evaluations given. The variable given the most attention is the type of job for which the candidates are considered. Typically, a prediction is made that females will be given lower evaluations compared with males when being considered for jobs assumed to be masculine in nature, that is, jobs that are either pre-

dominantly held by males (Muchinsky & Harris, 1977) or jobs that reflect demanding or challenging activities (Rosen & Jerdee, 1974b). What is proposed is a sex-congruency model whereby a situation or job is sex typed as being more appropriate for a male or a female and thereby influences the evaluation given.

To date the research provides fairly consistent evidence confirming the sex-congruency notion. Studies by Shaw (1972), Cohen and Bunker (1975), Cash, Gillen, and Burns (1977), and Rosen and Jerdee (1974b) show significant interactions between sex and type of job in their influence on evaluations. However, the study by Muchinsky and Harris (1977), in which individuals evaluated male and female resumes for jobs in mechanical engineering (traditionally masculine), a child-care center (traditionally feminine), and journalism (a neuter job), indicated only mixed support for the hypothesis of differential evaluations as a function of job type and sex. In different contexts, Feather (1975) and Feather and Simon (1975) found evidence for the differential evaluation of males and females as a function of the appropriateness of the occupation for one or the other sex.

An additional job variable that has been investigated is the predominant sex of the subordinates in the job. Rose and Andiappan (1978) hypothesized that greater differential evaluations of male and female job candidates would occur when interviewers evaluated such candidates for managerial jobs that involved either a predominantly male or a predominantly female work force. Male ($n = 55$) and female ($n = 20$) college students evaluated resumes for a managerial position. Sex of subject, sex of applicant, and predominant sex of subordinates were the variables investigated. The results of their resume study indicated that female candidates were evaluated more favorably when the predominant sex of the subordinate employees was female. Likewise, male candidates were given higher evaluations when their potential subordinate work force was predominantly male.

A further variable that has been investigated in several studies is applicant qualifications or competence. Two basic questions were behind the investigations of the effects of applicant qualifications. First, researchers were interested in whether significant sex effects emerged simply because sex was the only salient cue in the stimulus set and whether when sex was considered jointly with qualifications the sex effects would diminish or even disappear because of the powerful effect of the qualification variable. A second question concerned the possible interaction between applicant sex and qualifications and whether interviewers evaluate highly competent females more poorly than highly competent males in comparison with the differences in evaluations that occur when the candidates are not so well qualified (Spence & Helmreich, 1972). These predictions were based on the notion that highly competent or qualified women are particularly threatening to male interviewers. Several studies bear directly on these issues. Dipboye et al. (1975) found a large main effect for competence: Highly qualified job candidates were preferred over less qualified candidates. Although the sex of the applicant was still a significant main effect, it accounted for less than 1% of the variance in the ratings compared with 38% accounted for by the qualification variable. On the other hand, when subjects were asked to choose only one candidate from the resumes, 72% of them chose a male. Finally, there was no indication of an interaction between applicant qualifications and applicant sex.

Similar results were obtained by Dipboye et al. (1977). Although applicant qualifications accounted for a large portion of the evaluation variance compared with that accounted for by applicant sex (50% vs. .006%), when subjects were asked to choose only one candidate for hire, highly qualified males were selected by 54% of the subjects, whereas 28% chose the highly qualified females. This difference was significant at the .05 level.

The study reported by Muchinsky and Harris (1977) indicated that significant differences were found between males and females of average ability (as indicated by

scholastic standing) but that no differences occurred between male and female applicants at high or low ability levels.

Heneman (1977) reported a study in which 144 college students evaluated hypothetical applicants for the job of life insurance agent. The qualifications of applicants were manipulated indirectly through the use of test scores. In this study, highly qualified females were evaluated as being much less suitable for hire than were highly qualified males.

Finally, the resume study conducted by Haefner (1977) found that while highly competent individuals were preferred over less competent candidates, preferences were clearly given to highly competent males over highly competent females. Two research trends are suggested by these data. First, although qualifications of candidates are clearly the most powerful factors in influencing interviewers' decisions, sex of applicant remains a significant variable. Moreover, although the sex of the applicant does not appear to account for a great deal of rating variability, when only one or two hiring choices are given to evaluators, sex of the applicant is a highly important variable. This situation is analogous to the utility of a test that has low validity but is used in a situation in which a large number of candidates are considered and only a small number of jobs are available. The low selection ratio gives a test with low validity a high degree of utility (Wiggins, 1973).

The findings regarding the predicted interaction of applicant qualifications with rater sex are mixed. Some studies demonstrate support for this notion and others do not, or they indicate interactions other than those predicted.

An additional variable that was investigated in three studies is the attractiveness of the job candidates. Dipboye et al. (1975) investigated the notion that physical attractiveness may be a more important variable in influencing interview evaluations for females than for males. However, their results showed that attractive candidates were preferred over unattractive candidates regardless of sex. A second study by Dipboye et al. (1977) showed different results. Attractive males were rated significantly higher than were attractive females, and unattractive males were rated higher than were unattractive females, but no difference occurred between male and female candidates who were moderately attractive.

The third study investigating this variable was more complex. Cash et al. (1977) made several predictions:

1. When candidates are under consideration for neuter jobs (i.e., jobs that are not considered traditionally male or female, attractive applicants will be more favorably evaluated than unattractive candidates, regardless of sex.

2. When candidates are considered for a traditionally masculine job, attractive males will be more highly evaluated than attractive females. In contrast, when under consideration for a traditionally feminine job, attractive females will be more positively evaluated than attractive males.

Results of the study indicated that these predictions were for the most part confirmed. In summary, attractiveness is an important factor for both males and females, but may influence interviewers' decisions differently for males and females, again depending on the type of job under consideration.

Applicant Race

Somewhat surprisingly, only three studies have dealt with race of the applicant in interview evaluations. Even more surprising, however, is that little evidence was found to suggest that interviewers give more unfavorable evaluations to black job candidates compared with white candidates.

Wexley and Nemeroff (1974) conducted a study in which resumes of blacks or whites were presented to 120 students who were asked to evaluate either black or white candidates with regard to their suitability for the position of mechanical-engineer technician. In addition, each resume contained information indicating whether the candidate was relatively similar in background to the interviewer (e.g., father was an office worker and mother a school teacher) or dissimilar in background to the interviewer (e.g., father was a laborer and mother a domestic work-

er). The students were also divided into relatively high and low racially prejudiced groups on the basis of their scores on an attitude measure of prejudice. The results indicated that the race of the applicant had no effect on the employability ratings given. However, similarity of biographical background did prove to be a major determinant of the evaluations of job candidates. Moreover, interviewers who were relatively more prejudiced tended to give lower ratings to both white and black applicants than did interviewers who were relatively less prejudiced.

Haefner (1977) also used a resume design and found that race interacted with both sex and age to influence interviewers' evaluations, but these effects were very small.

Finally, Rand and Wexley (1975) showed 80 white males and 80 white females videotapes of employment interviews in which a black or white applicant was presented. Again, race of the applicant did not significantly affect the evaluations of the candidates. Additional results showed that biographical similarity and racial prejudice on the part of interviewers influenced the evaluations given, but these factors operated to influence the ratings of both black and white applicants. Thus, not a great deal of research evidence is available that demonstrates that interviewers give lower evaluations to black applicants who have qualifications equal to those of white applicants.

A factor that might account for the paucity of significant findings is that the studies were conducted within the past 5 years and that the interviewers generally consisted of students who may have been somewhat liberal or sensitive to the EEOC and legal issues associated with evaluating black applicants for jobs. They might have been apt to respond somewhat differently if they had not been sensitized to these issues. A further factor is that all the studies except that reported by Haefner (1977) used between-subjects designs in which subjects viewed only one stimulus condition (either the black or white applicant). Within-subject designs have two advantages that might be capitalized on in future studies: First, given the same total number of subjects, there are more degrees of freedom, thus making an experiment relatively more powerful in detecting differential evaluations if they, in fact, occur; second, subjects are allowed to view more than one stimulus condition, which allows them to establish anchor and reference points on the rating scales. The between-subjects designs in the studies reported above may have simply not been powerful enough to detect significant effects.

Applicant Age

Only two studies have investigated the effects of candidates' age on interviewers' evaluations. Haefner (1977) found that age was a significant factor in interviewers' evaluations and that age also interacted with race (as noted above) and with competence. Although age played no role in the evaluation of barely competent candidates, younger highly competent individuals were preferred over relatively older but highly competent individuals.

Using the in-basket methodology, Rosen and Jerdee (1976b) found some interesting results. Undergraduates ($N = 142$) reacted to six items embedded in an in-basket in which age was the variable manipulated. Older employees were evaluated as less suitable for a job compared with younger employees. In addition, the older employees were evaluated as (a) less promotable, (b) more resistant to change, (c) having lower physical capability, and (d) less likely to have organizational support for retraining opportunities. Although these two studies suggest a rather strong and pervasive effect due to applicant age, more research is needed to substantiate these findings.

Applicant Handicap

There is a relative dearth of studies demonstrating the effects (or lack of effects) of handicapped status on interview evaluations. An early study by Shaw (1972), for example, investigated the differential evaluation of job candidates when one was depicted as an individual with a "withered arm" and weak vision. Subjects were 132 college recruiters. Although no significant effects were

observed, the results indicated that the candidate with the physical disability was perceived relatively favorably. The authors' explanation was that subjects perceived the candidate "as a courageous figure who had overcome physical adversity, rather than an employment risk because of physical handicaps" (Shaw, 1972, p. 337).

Krefting and Brief (1977) reported the results of a study in which 145 college students evaluated a packet of resumes and other materials necessary for determining whether an applicant should be hired for the position of typist. The applicant in the resume was depicted as either healthy or confined to a wheelchair. Also, the applicant was depicted as either experienced or nonexperienced.

The individuals reviewing the resume materials gave estimates of the applicant's health, motivation, potential for staying, and so forth in addition to an overall rating. Disabled applicants were seen as significantly more highly motivated and as more likely to become long-term employees than were nondisabled applicants. However, a puzzling interaction also occurred. The inexperienced, disabled applicant was evaluated higher overall than was the experienced, disabled applicant. The authors suggested that the results of the study were relatively encouraging for qualified, disabled applicants. One of the problems with this and similar studies is that the subjects may have guessed the purpose of the study because of the nature of the applicant variable (i.e., handicapped status) displayed.

In a second study along these lines, Rose and Brief (Note 1) assessed the impact of applicant disability (epilepsy) and job type on evaluation judgments. Students ($N = 145$) evaluated a resume and other data that portrayed the applicant as either healthy or having epilepsy (although the seizures were under control). Applicants were portrayed as applying for a job that involved either a great deal of public contact and supervisory responsibilities or no public contact or supervisory responsibilities. Subjects gave evaluations along several dimensions. Results indicated that epileptic and normal applicants were not perceived to be significantly different in terms of their overall performance, tenure with the firm, work effectiveness, and amount of salary. There was, however, a significant interaction between disability and type of job on several evaluations. Epileptics seeking non-public-contact jobs were perceived as more likely to satisfy clients than were normal applicants.

A third study is perhaps not so encouraging. Johnson and Heal (1976) reported the results of a study that again compared a wheelchair job applicant with a healthy applicant. In this study, actual employment agencies (50) were visited by one of the researchers, who indicated interest in finding a job. However, in half the interviews, the same researcher appeared in a wheelchair. The researcher rated the agency responses on a number of variables. The analyses of the data showed that handicapped applicants as opposed to healthy applicants were offered significantly fewer future job interviews and given a more gloomy estimation of the job market by the agency representatives. In addition, the representatives gave the handicapped applicants a lower probability of getting the kind of work that was being sought and offered relative discouragement to the handicapped applicants about seeking a position with normal public exposure.

The results of this study may have been influenced by the fact that the researcher did the ratings herself. However, the findings certainly suggest that interviewers may give less favorable evaluations to individuals displaying a handicap.

Summary of Evidence Concerning Differential Evaluations

A review of the research concerning differential evaluations of minority job candidates based on components of the interview yields mixed results. First, the evidence based on the resume research is fairly consistent in showing that women tend to be evaluated more poorly than men. Moreover, the degree of differential evaluation appears to be related to the type of job for which women are considered; a more prominent bias occurs when women are considered for typically masculine-oriented jobs.

When qualifications of candidates are considered, research studies tend to demonstrate the following: (a) Qualifications of job candidates show a powerful main effect that accounts for 25%–50% of the variance in ratings; (b) the predicted interaction between applicant qualifications and sex was not consistently found and, in particular, did not support the notion that highly competent women are prone to more negative evaluations compared with highly competent males in relation to differences between less qualified male and female applicants.

A somewhat surprising finding is the paucity of evidence for differential evaluation of job candidates as a function of race. Although only a few studies have investigated this phenomenon, the data do not support typical a priori assumptions about the interview's providing a ready mechanism by which to discriminate against blacks.

This review also reveals a relative dearth of research investigating the possible bias in interview circumstances that operates to negatively affect elderly applicants as well as those with handicaps. Although the existing studies suggest a strong and pervasive effect due to applicant age, the data dealing with handicapped applicants are more complex. The few studies that have explored this issue suggest that handicapped applicants are viewed as having higher motivational tendencies, possibly because of the perceptions of interviewers that handicapped applicants must have exerted a great deal of effort, worked harder, and so on in order to overcome their disabilities. Obviously, more research needs to be performed on the kinds of evaluations handicapped individuals receive in employment settings.

Differential Validity of the Interview

A major thrust of the 1970 EEOC guidelines on testing was the need for employers to investigate the possibility that a selection device may exhibit differential validity. Boehm (1972) and Humphreys (1973) have indicated that differential validity has to do with whether the correlation between predictor and criterion for one subgroup of a population differs significantly from the correlation between a predictor and criterion for a different subgroup of a population.

Although numerous studies have examined the validity of the interview and interview process in general (Mayfield, 1964; Ulrich & Trumbo, 1965; Wright, 1969), only three studies were located that examined the validity of the interview as a function of employee subgroup characteristics.

Lopez (1966) reported the results of a study in which black and white female applicants for toll collection jobs were given a standardized 10-minute biographical data interview in which training, experience, and personal qualifications were numerically assessed. Although the study was seriously flawed, the data indicated that the interview was significantly correlated with toll accuracy and length of time in service for whites. For the black sample, the interview was significant when correlated against the criteria of attendance, toll accuracy, length of service, and supervisor's appraisal. However, these results are extremely suspect because the correlations were corrected for restriction of range, sample size values were not reported for the separate groups, and tests of differential validity were not conducted.

Kirkpatrick, Ewen, Barrett, and Katzell (1968) reported validity coefficients separately by race. Interviewers' ratings based on a 5-point scale were correlated against six different criterion measures for 94 white and 22 black female clerical workers. The interview ratings lacked validity in the total sample and in each group separately. Thus, the differential validity hypothesis was not supported;[15] judgments based on the interview were invalid for both groups. In addition, no difference was observed between the ratings given to the black and white groups by interviewers.

Freytag (Note 2) reported the results of a study designed to investigate the differential validity of the interview in predicting performance of 64 minority members (blacks, females, Hispanics, etc.) and 54 white pa-

[15] These authors tested for differential validity incorrectly, but since both coefficients did not differ significantly from zero, one can infer that no differential validity occurred in this instance.

trol officers. Interview ratings based on a three-man oral board were correlated with performance ratings. Only interview ratings of the officers' communication skills correlated significantly for the total sample; however, a test of differential validity indicated that the interview was equally valid for both groups.

To summarize these three studies, very little evidence exists that indicates that the interview is differentially valid for minority group members and nonminority members. On the other hand, little evidence indicates that the phenomenon does not exist because of the paucity of research investigating this issue.

Concluding Comments and Needed Research

For a selection device that is so widely used, we know very little about its discriminatory impact on minorities. Given the high probability that the interview will be subject to greater judicial review, it is somewhat startling to realize that only a little over 20 studies have investigated the issue of differential evaluations or validity of the interview.

Throughout this article, a number of research needs have been identified. Some of these are described below.

Methodological Needs

More research is needed that uses different methodologies in investigating possible differential evaluation or bias in the interview. Presently, there seems to be an overreliance on resume techniques. It seems likely that evaluations given to "pencil-and-paper" people differ from those given to more fully described and portrayed stimulus people. More efforts should be made to study in vivo interview simulations or to use methodologies that present fuller stimulus fields of the interview situation. The studies by Wexley and Nemeroff (1974), Wexley, Sanders, and Yukl (1973), and Dipboye and Wiley (1977) are excellent examples of the presentation of stimulus material through videotape techniques. In addition, more efforts to conduct field research using the methodology suggested by Johnson and Heal (1976) should be made. Face-to-face interviews also need to be studied.

Future researchers may also want to explore some Bayesian methodologies, concepts, and approaches (Hayes, 1973; Slovic & Lichtenstein, 1971) in consideration of the differential evaluation issue. Such techniques might be used to examine the revision of interview judgments or prior probabilities of interviewers' success as more information about candidates becomes available. Prior probabilities of success in jobs for minorities may be more resistant to change than prior probabilities for nonminorities. Moreover, the prior-probability distributions may themselves be different based on the minority/majority status of the interviewee. Osburn and Constantin (Note 3) suggested that interviewers, when lacking information on bona fide qualifications, respond as if women have a lower probability of success in jobs traditionally held by women. Thus, the formation of differential evaluations of candidates might be viewed from a Bayesian perspective.

Between-subjects and within-subject designs need to be contrasted with regard to the information they yield regarding differential evaluations. Within-subject designs may be more powerful in detecting differential evaluation for two reasons: First, from a statistical perspective, more degrees of freedom are available. Second, interviewers themselves may be more prone to give differential evaluations when they have comparative stimulus sets of interviewees.

Researchers also need to tap subject pools other than undergraduate and graduate students. Though some research indicates that college student reactions are not appreciably different from those of managers (Bernstein, Hakel, & Harlan, 1975; Dipboye et al., 1975), more real-life interviewers need to be used in research of this sort.

Finally, researchers should be more aware of the need to present greater stimulus sampling in the conditions presented; that is, subjects should have the opportunity to view more than one particular female or handicapped individual in studies of this kind. Otherwise, any significant effect observed

might be unique to the specific stimulus individual presented because of other uncontrolled minor characteristics (hair color, height, etc.). Pools of stimulus applicants from which samples can be drawn to show subjects should be developed.

Research on Race, Age, and the Handicapped

Research is needed to further establish whether interview judgments are biased against the elderly, the handicapped, or other minorities. Little interview research has been conducted on these minorities. Given the relatively high probability of more lawsuits' occurring in these areas, it would be wise to focus greater research efforts on studying the interview as it may impact on these particular groups. For example, an area of research that demands further exploration is the notion that subjects overevaluate handicapped applicants (Krefting & Brief, 1977). Perhaps in these situations subjects attribute higher motivation, ability levels, and so forth to handicapped individuals filing applications for jobs because of their perceived efforts to overcome their disabilities.

Further research might also focus on differential evaluations as a function of type of disability. Some disabilities (e.g., mental illness) are perceived more negatively by employers than are other disabilities (Nagi, McBroom, & Colletts, 1972). Thus, lower evaluations may be given to applicants with the most negatively viewed disability or handicap.

Similarly, research is needed to investigate a possible Age × Type of Job interaction. It seems likely that there are some jobs in which older individuals are perceived as the more desirable candidates (e.g., president of company, chairperson of board) and some jobs in which old age is perceived as a handicap (e.g., airplane pilot). The possible interactions need to be explored.

Process Research

More research is needed to determine what goes on in interviews to influence differential evaluations; that is, researchers should begin to focus on the underlying process by which differential evaluations take place. As noted earlier, the exact methods by which stereotypes affect interviewer judgments are not known. Do interviewers base their differential judgments on the matching or the stereotyping paradigms described earlier?

Many of the questions asked earlier about the process of interviewer judgments (Webster, 1964) need to be further examined in the context of differential evaluations. For example, a number of process issues and areas need further investigation. When do differential evaluations occur during the interview process—early or late? Do interviewers gather different kinds of information or pay attention to different information depending on the majority or minority status of the candidate? Policy-capturing methodologies along the lines of those employed by Zedeck and Kafry (1977) might be useful in these contexts.

What kinds of differences occur between studies that examine the impact of face-to-face contact with applicants and studies that assess candidates on the basis of resumes? Imada and Hakel (1977) showed that nonverbal communication (e.g., eye contact, posture, gesturing, etc.) has a strong impact on interviewer decisions. These behavioral components of the interview process need to be examined with regard to their impact on differential evaluations. In short, more molecular studies of the verbal and nonverbal aspects of the interview need to be carried out.

Does the amount of information about candidates and jobs affect differential evaluations? Studies by Longsdale and Weitz (1973) indicate that interviewers are more prone to agree about job candidates when there is a great deal of information concerning the job than when there is not much information about the job. Wiener and Schneiderman (1974) conducted two experiments showing that complete and unambiguous job information reduces the effect of irrelevant stereotypes on interviewer decisions. Does this also result in a minimization of differential evaluations? Similarly, does providing more information about the applicant also serve to decrease differential evaluations?

Strength of Effects

Analyses that yield precise estimates of the proportion of variance accounted for by the minority/majority status of the candidates in interview situations need to be carried out. The few studies that have calculated and presented such estimates indicate that minority status accounts for relatively small amounts of the variation. Yet, such effects may have decidedly powerful effects when only a small number (e.g., one) of the candidates are chosen as the best or most likely candidates.

Differential Validity and Regression

More research is needed to investigate the possibilities of differential regression in the interview. Is the interview equally valid for blacks and whites, women as well as men, and so on? It may be that if interviewers collect and process different information in an interview based on the minority/majority status of job candidates, then the interview may indeed show different patterns of validity with different job criterion measures. Similarly, more psychometrically sophisticated models should be used in reviewing the fairness of the interview.

Interviewer Training

Does interviewer training affect the kinds of evaluations given to minority and majority job candidates? Studies by Wexley et al. (1973) and Latham, Wexley, and Purcell (1975) have shown that workshop training of interviewers geared toward the reduction of rating errors is fairly effective. Do efforts to provide interviewers with training about how to interview minorities, what questions to avoid, and so forth result in more accurate evaluations and the reduction of differential evaluation? To date no studies exist concerning these issues.

Interviewer Variables

Additional efforts should be made to explore possible factors associated with the interviewer that may affect any differential evaluations given. Arvey, Passino, and Lounsbury (1977) found that sex of the evaluator showed marginal effects on the gathering of job analysis data. London and Poplawski (1976) reported that females tend to make more favorable evaluations than do males. In addition, Rumenik, Capasso, and Hendrick (1977) summarized a number of studies concerning sex of the interviewer and concluded that this variable is indeed a potent one.

Moreover, other interviewer characteristics should be investigated with regard to their possible influence on differential evaluations. For example, Rose and Andiappan (1978) suggested that the interviewer differences in androgyny may influence differential evaluation.

Integration of Findings With Other Research

Research should be conducted that resolves the conflict between many of the findings gathered in the context of interview research and those gathered in other contexts. For example, Hamner, Kim, Baird, and Bigoness (1974), Bigoness (1976), and Jacobson and Effertz (1974), among others, have found that females are evaluated more favorably when raters evaluate performance. Why the results from these studies differ substantially from those summarized earlier needs to be determined.

It is hoped that the present article will serve to direct and foster future research efforts. The last word about the interview is a long way from being said.

Reference Notes

1. Rose, G. L., & Brief, A. *The impact of job type and applicant disability on judgments in the selection process.* Paper presented at the meeting of the Academy of Management, Orlando, Fla., August 1977.
2. Freytag, W. R. *The validity of the oral board interview in police officer selection: A comparative analysis between minority-group members and white males.* Unpublished manuscript, Pennsylvania State University, Applied Research Laboratory, 1976.
3. Osburn, H. G., & Constantin, S. W. *The selection interview: Some issues raised by analogue research on decision-making.* Unpublished manuscript, University of Houston, 1977.

References

Arvey, R. D. *Fairness in selecting employees.* Reading, Mass.: Addison-Wesley, 1979.

Arvey, R. D., Passino, E. M., & Lounsbury, J. N. Job analysis results as influenced by sex of incumbent and sex of analyst. *Journal of Applied Psychology,* 1977, *62,* 411–416.

Babcock, B., Freedman, A., Norton, E., & Ross, A. *Sex discrimination and the law: Causes and remedies.* Boston: Little, Brown, 1975.

Bernstein, V., Hakel, M. D., & Harlan, A. The college student as interviewer: A threat to generalizability? *Journal of Applied Psychology,* 1975, *60,* 266–268.

Bigoness, W. J. Effect of applicant's sex, race, and performance on employer's performance ratings: Some additional findings. *Journal of Applied Psychology,* 1976, *61,* 80–84.

Boehm, V. R. Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology,* 1972, *56,* 33–39.

Brigham, J. C. Ethnic stereotypes. *Psychological Bulletin,* 1971, *76,* 15–38.

Britton, J. O., & Thomas, K. R. Age and sex as employment variables: Views of employment service interviewers. *Journal of Employment Counseling,* 1973, *10,* 180–186.

Cash, T. F., Gillen, B., & Burns, D. S. Sexism and "beautyism" in personnel consultant decision making. *Journal of Applied Psychology,* 1977, *62,* 301–307.

Cecil, E. A., Paul, R. J., & Olins, R. A. Perceived importance of selected variables used to evaluate male and female job applicants. *Personnel Psychology,* 1973, *26,* 397–404.

Cleary, T. A. Test bias: Predictions of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement,* 1968, *5,* 115–124.

Cohen, S. L., & Bunker, K. A. Subtle effects of sex role stereotypes on recruiters' hiring decisions. *Journal of Applied Psychology,* 1975, *60,* 566–572.

Cole, N. S. Bias in selection. *Journal of Educational Measurement,* 1973, *10,* 237–255.

Dipboye, R. L., Arvey, R. D., & Terpstra, D. E. Sex and physical attractiveness of raters and applicants as determinants of resumé evaluations. *Journal of Applied Psychology,* 1977, *62,* 288–294.

Dipboye, R. L., Fromkin, H. L., & Wiback, K. Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant résumés. *Journal of Applied Psychology,* 1975, *60,* 39–43.

Dipboye, R. L., & Wiley, J. W. Reactions of college recruiters to interviewer sex and self-presentation style. *Journal of Vocational Behavior,* 1977, *10,* 1–12.

Dunnette, M. D., & Bass, B. M. Behavioral scientists and personnel management. *Industrial Relations,* 1963, *2,* 115–130.

Equal Employment Opportunity Commission. Guidelines on employee selection procedures. *Federal Register,* 1970, *35,* 12333–12336.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register,* 1978, *43,* 38290–38315.

Feather, N. T. Positive and negative reactions to male and female success and failure in relation to the perceived status and sex-typed appropriateness of occupations. *Journal of Personality and Social Psychology,* 1975, *31,* 536–548.

Feather, N. T., & Simon, J. G. Reactions to male and female success and failure in sex-linked occupations: Impressions of personality, causal attributions, and perceived likelihood of different consequences. *Journal of Personality and Social Psychology,* 1975, *31,* 20–31.

Fidell, L. S. Empirical verification of sex discrimination in hiring practices in psychology. *American Psychologist,* 1970, *25,* 1094–1098.

Fugita, S. S., Wexley, K. N., & Hillery, J. M. Black-white differences in nonverbal behavior in an interview setting. *Journal of Applied Social Psychology,* 1974, *4,* 343–350.

Haefner, J. E. Race, age, sex, and competence as factors in employer selection of the disadvantaged. *Journal of Applied Psychology,* 1977, *62,* 199–202.

Hall, E. T. *The hidden dimensions.* New York: Doubleday, 1966.

Hamilton, D. L. Cognitive biases in the perception of social groups. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior.* Hillsdale, N.J.: Erlbaum, 1976.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology,* 1974, *59,* 497–499.

Hayes, W. L. *Statistics for the social sciences.* New York: Holt, Rinehart & Winston, 1973.

Heneman, H. G., III. Impact of test information and applicant sex. *Journal of Applied Psychology,* 1977, *62,* 524–526.

Humphreys, L. G. Statistical definitions of test validity for minority groups. *Journal of Applied Psychology,* 1973, *58,* 1–14.

Hunter, J. E., & Schmidt, F. L. Critical analysis of the statistical and ethical implications of various definitions of *test bias. Psychological Bulletin,* 1976, *83,* 1053–1071.

Imada, A. S., & Hakel, M. D. Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology,* 1977, *62,* 295–300.

Jacobson, M. B., & Effertz, J. Sex roles and leadership: Perceptions of the leaders and the led. *Organizational Behavior and Human Performance,* 1974, *12,* 383–396.

Johnson, R., & Heal, L. W. Private employment agency responses to the physically handicapped applicant in a wheelchair. *Journal of Applied Rehabilitation Counseling,* 1976, *7,* 12–21.

Jones, J. M. *Prejudice and racism.* Reading, Mass.: Addison-Wesley, 1972.

Karlins, M., Coffman, T. L., & Walters, G. On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology,* 1969, *13,* 1–16.

Katz, D. The functional approach to the study of attitudes. *Public Opinion Quarterly,* 1960, *24,* 163–177.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. *Testing and fair employment: Fairness and validity of personnel tests for different ethnic groups.* New York: New York University Press, 1968.

Krefting, L. A., & Brief, A. P. The impact of applicant disability on evaluative judgments in the selection process. *Academy of Management Journal,* 1977, *19,* 675–680.

Kryger, B. R., & Shikiar, R. Sexual discrimination in the use of letters of recommendation: A case of reverse discrimination. *Journal of Applied Psychology,* 1978, *63,* 309–314.

Latham, G. P., Wexley, K. M., & Purcell, E. D. Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology,* 1975, *60,* 550–555.

Lippman, W. *Public opinion.* New York: Harcourt, Brace, 1922.

London, M., & Poplawski, J. R. Effects of information on stereotype development in performance appraisal and interview contexts. *Journal of Applied Psychology,* 1976, *61,* 199–205.

Longsdale, J., & Weitz, J. Estimating the influence of job information on interviewer agreement. *Journal of Applied Psychology,* 1973, *57,* 23–27.

Lopez, F. M., Jr. Current problems in test performance of job applicants. *Personnel Psychology,* 1966, *19,* 10–17.

Mayfield, E. C. The selection interview: A reevaluation of published research. *Personnel Psychology,* 1964, *17,* 239–260.

Muchinsky, P. M., & Harris, S. L. The effect of applicant sex and scholastic standing on the evaluation of job applicant resumes in sex-typed occupations. *Journal of Vocational Behavior,* 1977, *11,* 95–108.

Nagi, S., McBroom, W. H., & Colletts, J. Work, employment and the disabled. *American Journal of Economics and Society,* 1972, *31,* 20–34.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement,* 1976, *13,* 3–29.

Rand, T. M., & Wexley, K. N. Demonstration of the effect, "similar to me," in simulated employment interviews. *Psychological Reports,* 1975, *36,* 535–544.

Renwick, P. A., & Tosi, T. The effects of sex, marital status and educational background on selection decisions. *Academy of Management Journal,* 1978, *21,* 93–103.

Rose, G. L., & Andiappan, P. Sex effects on managerial hiring decisions. *Academy of Management Journal,* 1978, *21,* 104–112.

Rosen, B., & Jerdee, T. H. Effects of applicant's sex and difficulty of job on evaluations of candidates for management positions. *Journal of Applied Psychology,* 1974, *59,* 511–512. (a)

Rosen, B., & Jerdee, T. H. Influence of sex role stereotypes on personnel decisions. *Journal of Applied Psychology,* 1974, *59,* 9–14. (b)

Rosen, B., & Jerdee, T. H. *Becoming aware.* Chicago: Science Research Associates, 1976. (a)

Rosen, B., & Jerdee, T. H. The influence of age stereotypes on managerial decisions. *Journal of Applied Psychology,* 1976, *61,* 428–432. (b)

Rosen, B., & Jerdee, T. H. The nature of job-related age stereotypes. *Journal of Applied Psychology,* 1976, *61,* 180–183. (c)

Rumenik, D. K., Capasso, D. R., & Hendrick, C. Experimenter sex effects in behavioral research. *Psychological Bulletin,* 1977, *84,* 852–877.

Schaie, K. W. Translations in gerontology—From lab to life: Intellectual functioning. *American Psychologist,* 1974, *29,* 802–807.

Schein, V. E. The relationship between sex role stereotypes and requisite management characteristics. *Journal of Applied Psychology,* 1973, *57,* 95–100.

Schein, V. E. Relationships between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology,* 1975, *60,* 340–344.

Schmidt, N. Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology,* 1976, *29,* 79–102.

Shaw, E. A. Differential impact of negative stereotyping in employee selection. *Personnel Selection,* 1972, *25,* 333–338.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* 1971, *6,* 649–744.

Spence, J. T., & Helmreich, R. Who likes competent women? Competence, sex role congruence of interests, and subjects' attitudes toward women as determinants of interpersonal attraction. *Journal of Applied Social Psychology,* 1972, *2,* 197–213.

Sterrett, J. H. The job interview: Body language and perceptions of potential effectiveness. *Journal of Applied Psychology,* 1978, *63,* 388–390.

Terborg, J. R. Women in management: A research review. *Journal of Applied Psychology,* 1977, *62,* 647–664.

Terborg, J. R., & Ilgen, D. R. A theoretical approach to sex discrimination in traditionally masculine organizations. *Organizational Behavior and Human Performance,* 1975, *13,* 352–376.

Thorndike, R. L. Concepts of culture-fairness. *Journal of Educational Measurement*, 1971, *8*, 63–70.
Ulrich, L., & Trumbo, D. The selection interview since 1949. *Psychological Bulletin*, 1965, *63*, 100–116.
Washington State Human Rights Commission. State & local laws: Text of laws. *Fair Employment Practices*, 1979, Sec. 457, 2923–2926.
Webster, E. C. (Ed.). *Decision making in the employment interview*. Montreal, Canada: Eagle, 1964.
Wexley, K. N., & Nemeroff, W. F. The effects of racial prejudice, race of applicant, and biographical similarity on interviewer evaluations of job applicants. *Journal of Social and Behavioral Sciences*, 1974, *20*, 66–78.
Wexley, K. N., Sanders, R. E., & Yukl, G. A. Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology*, 1973, *57*, 233–236.
Wiener, Y., & Schneiderman, M. L. Use of job information as a criterion in employment decisions of interviewers. *Journal of Applied Psychology*, 1974, *59*, 699–704.
Wiggins, J. S. *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley, 1973.
Wright, O. R., Jr. Summary of research on the selection interview since 1969. *Personnel Psychology*, 1969, *22*, 391–413.
Zedeck, S., & Kafry, D. Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance*, 1977, *18*, 269–294.
Zikmund, W. G., Hitt, M. A., & Pickens, B. A. Influence of sex and scholastic performance on reactions to job applicant resumes. *Journal of Applied Psychology*, 1978, *63*, 252–255.

Received February 28, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 766–776

Linear Models for the Analysis and Construction of Instruments in a Facet Design

Gideon J. Mellenbergh, Henk Kelderman, Jenneke G. Stijlen, and Edu Zondag
University of Amsterdam, Amsterdam, The Netherlands

Linear models are described for the situation wherein a measurement instrument is constructed for all elements of the Cartesian product of several facets when the elements of each facet are not ordered. The structure of the covariance matrix of the instrument is derived from the models. By using covariance structure analysis, the models can be tested, and estimates of the parameters can be obtained. Models for 20 tests were formulated and tested and were constructed from a design with a behavioral and a situational facet measuring social anxiety in children; for 15 tests a model proved to fit the data. It is concluded that covariance structure analysis is useful for the analysis and construction of measurement instruments.

Facet designs can be used for test item writing. A facet is a set consisting of a finite number of elements (Foa, 1965). The Cartesian product of a finite number of facets is a design for item and test construction. This design for test construction corresponds to the factorial design for experimentation. Facet designs have been used in different areas of psychological and educational measurement. For example, Guttman and Schlesinger (1967) constructed distractors for ability and achievement items from facet designs. Hamersma, Paige, and Jordan (1973) reported facet designs for the construction of attitude scales. Butt and Fiske (1968) compared dominance scales constructed from a facet design with dominance scales developed using other strategies. Guilford's (1967) structure-of-intellect model can be conceived as a facet design (Fiske, 1971, p. 128). The multitrait–multimethod matrix (Campbell & Fiske, 1959) results from a facet design: The correlation coefficients of this matrix are computed between the elements of the Cartesian product of a facet with traits and a facet with methods.

Scores from tests constructed in a facet design can be analyzed using generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). For example, Endler and Hunt (1966) used a facet design with 11 situations and 14 modes of response. For each combination of situation and mode of response a 5-point scale was constructed. By considering subjects as elements of a separate facet and by assuming that the elements of each facet are a random sample from a universe, variance components were estimated for the situations, the modes of response, the subjects, and their interactions. Scores from tests constructed in a facet design can also be analyzed using three-mode factor analysis (Tucker, 1963). For example, Levin (1965) analyzed the Endler and Hunt data with this technique. The results of this analysis are three matrices: one with factor loadings of the situations, one with factor loadings with the responses, and a core. The core contains the scores for the interaction of situations and responses for types of subject.

Another approach is the analysis of measures of association between the instruments constructed in a facet design. For example, Schlesinger and Guttman (1969) computed correlation coefficients between intelligence

We thank David A. Kenny, Fred N. Kerlinger, Wim J. van der Linden, Wim E. Saris, and Pieter Vijn for their comments and Ineke Wesling for typing the manuscript.

Requests for reprints should be sent to Gideon J. Mellenbergh, Subfaculty of Psychology, Department of Methodology, University of Amsterdam, Weesperplein 8, Amsterdam, The Netherlands.

and achievement tests in a facet design; they used smallest space analysis for the analysis of the correlation matrix. Bock and Bargmann (1966) analyzed covariance matrices of tests constructed in a facet design. They considered the elements of the facets from which the tests were constructed as fixed and the elements of the facet with subjects as random. Assuming that the observed scores were linear functions of a set of uncorrelated latent variables, they analyzed the covariance matrices and estimated the variances of the latent variables. Wiley, Schmidt, and Bramble (1973) generalized the models considered by Bock and Bargmann.

New models have recently been developed for the analysis of covariance matrices: Hypotheses about the structure of the matrix can be formulated and tested; estimates of the parameters in the hypothesized model can be computed (Bentler, 1976; Jöreskog, 1970, 1973; Long, 1976; Mukherjee, 1973). A facet design is preeminently suited to generating and formulating hypotheses about covariance matrices. Covariance structure analysis should therefore be considered as an important method for the analysis of covariance and correlation matrices computed between instruments constructed from a facet design; it can further an "integration of test design and analysis" (Guttman, 1970).

Linear models are formulated for facet designs. The models are applied to a correlation matrix of 20 instruments from a facet design measuring social anxiety in children. Using covariance structure analysis several models were tested; parameter estimates are reported for the model that best fits the data.

Models for Facet Designs

The situation wherein the elements of a facet are not ordered is considered; instruments are constructed for all elements of the Cartesian product of the facets. A general model for this situation is described; from this model special cases are derived. For the sake of simplicity, the models are described for two facets with p and q elements; the models can easily be generalized to more than two facets. The two facets can be represented by a rectangle with p rows and q columns. Each square of the rectangle represents an element of the Cartesian product of the two facets (for an example, see Table 1). It is assumed that for each element of the Cartesian product, one measurement instrument is constructed.

It is also assumed that the score of Subject i on the instrument from the jth row and the kth column is a linear function of a set of latent variables:

$$y_{ijk} = m_{jk} + s_{jk}S_i + r_{jk}R_{ij} + c_{jk}C_{ik} + rc_{jk}RC_{ijk} + E_{ijk}. \quad (1)$$

In this equation m_{jk} is the mean score on the instrument in the population of subjects; the parameters s_{jk}, r_{jk}, c_{jk}, and rc_{jk} are loadings specific to the jth row and the kth column of the facet design. The other terms in the equation represent the scores of Subject i on random latent variables: S_i is the score on a general latent variable, R_{ij} and C_{ik} are the scores on

Table 1
Cartesian Product of a Situational and Behavioral Facet for Social Anxiety in Children

Situational facet	Behavioral facet: Cognitive	Avoidance	Physiological	Affective
Social	Social–cognitive (y_{11})	Social–avoidance (y_{12})	Social–physiological (y_{13})	Social–affective (y_{14})
Intellectual	Intellectual–cognitive (y_{21})	Intellectual–avoidance (y_{22})	Intellectual–physiological (y_{23})	Intellectual–affective (y_{24})
Physical	Physical–cognitive (y_{31})	Physical–avoidance (y_{32})	Physical–physiological (y_{33})	Physical–affective (y_{34})
Exclusion	Exclusion–cognitive (y_{41})	Exclusion–avoidance (y_{42})	Exclusion–physiological (y_{43})	Exclusion–affective (y_{44})
Appearance	Appearance–cognitive (y_{51})	Appearance–avoidance (y_{52})	Appearance–physiological (y_{53})	Appearance–affective (y_{54})

latent variables specific to the jth row and the kth column, and RC_{ijk} is the score on a latent variable specific to the combination on the jth row and the kth column; E_{ijk} is the error score. The terms RC_{ijk} and E_{ijk} are both scores of Subject i on latent variables specific to the combination of the jth row and the kth column and cannot be separated without replications. Therefore, the term $rc_{jk}RC_{ijk}$ is absorbed into the residual term: $E_{ijk}^* = rc_{jk}RC_{ijk} + E_{ijk}$; omitting the asterisk in the residual yields the following model:

$$y_{ijk}=m_{jk}+s_{jk}S_i+r_{jk}R_{ij}+c_{jk}C_{ik}+E_{ijk}. \quad (2)$$

Restricting parameters in Equation 2 yields special cases. First, one or two of the parameters s_{jk}, r_{jk}, and c_{jk} are set equal to zero:

$$y_{ijk} = m_{jk} + r_{jk}R_{ij} + c_{jk}C_{ik} + E_{ijk}; \quad (3)$$

$$y_{ijk} = m_{jk} + s_{jk}S_i + c_{jk}C_{ik} + E_{ijk}; \quad (4)$$

$$y_{ijk} = m_{jk} + s_{jk}S_i + r_{jk}R_{ij} + E_{ijk}; \quad (5)$$

$$y_{ijk} = m_{jk} + s_{jk}S_i + E_{ijk}; \quad (6)$$

$$y_{ijk} = m_{jk} + r_{jk}R_{ij} + E_{ijk}; \quad (7)$$

$$y_{ijk} = m_{jk} + c_{jk}C_{ik} + E_{ijk}. \quad (8)$$

Second, the parameters can be set equal to one; Equation 2 reduces to

$$y_{ijk} = m_{jk} + S_i + R_{ij} + C_{ik} + E_{ijk}. \quad (9)$$

This model can be restricted, thus further eliminating one or two of the latent variables. Third, in Equations 2–9 parameters can be constrained in the sense that they are set equal to each other.

Equation 2 can also be formulated in vector notation:

$$y_{ijk} = m_{jk} + \mathbf{l}'_{jk}\mathbf{f}_i + E_{ijk}, \quad (10)$$

where $\mathbf{f}'_i$ is the vector with the scores of Subject i on all latent variables $[\mathbf{f}'_i = (S_i, R_{i1}, \ldots, R_{ij}, \ldots, R_{ip}, C_{i1}, \ldots, C_{ik}, \ldots, C_{iq})]$ and $\mathbf{l}'_{jk}$ is the vector with three parameters in positions corresponding to the positions of the elements S_i, R_{ij}, and C_{ik} in the vector $\mathbf{f}'_i$ and is zero otherwise $[\mathbf{l}'_{jk} = (s_{jk}, 0, \ldots, r_{jk}, 0, \ldots, c_{jk}, 0, \ldots)]$. The models displayed in Equations 3–8 can also be written in the form of Equation 10 by setting the appropriate elements of the vector $\mathbf{l}'_{jk}$ equal to zero; the model in Equation 9 is written in the form of Equation 10 by setting the three parameters of the vector $\mathbf{l}'_{jk}$ equal to one.

In Equation 10 the score of Subject i on one instrument is considered. However, for Subject i scores are obtained on all $p \times q$ instruments. The scores of Subject i on the instruments are collected in one observation vector: $\mathbf{y}'_i = (y_{i11}, \ldots, y_{i1q}, y_{i21}, \ldots, y_{i2q}, \ldots, y_{ip1}, \ldots, y_{ipq})$. From Equation 10 follows the model for the vector of observations of Subject i:

$$\mathbf{y}_i = \mathbf{m} + \mathbf{L}\mathbf{f}_i + \mathbf{e}_i, \quad (11)$$

where the vector $\mathbf{m}' = (m_{11}, \ldots, m_{1q}, m_{21}, \ldots, m_{2q}, \ldots, m_{p1}, \ldots, m_{pq})$ contains the population means of all instruments; the vector $\mathbf{e}'_i = (E_{i11}, \ldots, E_{i1q}, E_{i21}, \ldots, E_{i2q}, \ldots, E_{ip1}, \ldots, E_{ipq})$, the residuals of Subject i on all instruments; and $\mathbf{L}$ is the matrix with rows $\mathbf{l}'_{jk}$: $\mathbf{L}' = (\mathbf{l}_{11}, \ldots, \mathbf{l}_{1q}, \mathbf{l}_{21}, \ldots, \mathbf{l}_{2q}, \ldots, \mathbf{l}_{p1}, \ldots, \mathbf{l}_{pq})$. It is assumed without loss of generality that the latent variables have zero population means, with covariance matrix $\mathbf{E}^2$ for the residuals and $\mathbf{C}$ for the other latent variables. From the assumption that the population covariances between the residuals and the other latent variables are zero and from Equation 11, it follows that the population covariance matrix $\mathbf{S}$ of the observed scores on the instruments has the following structure (Jöreskog, 1974):

$$\mathbf{S} = \mathbf{LCL}' + \mathbf{E}^2. \quad (12)$$

The restrictions on the parameters mentioned above give rise to special cases of the matrix $\mathbf{L}$. Equation 6 implies that only the first column of $\mathbf{L}$ contains parameters; the other elements are zero. The model is a one-factor model; if this model fits the data, the instruments are termed *congeneric* tests (Jöreskog, 1971, 1974). The models displayed in Equations 3, 4, 5, 7, and 8 imply that the matrix $\mathbf{L}$ contains columns with both parameters and zeros; the models are restricted factor models (Lawley & Maxwell, 1971, chap. 7). Equation 9 implies that $\mathbf{L}$ is a matrix with zeros and ones, which in experimental design models is termed a *design* or *incidence* matrix (Searle, 1971, p. 166); the model is a variance component model (Jöreskog, 1974).

Assumptions about the covariances between the latent variables yield special cases of the matrices $\mathbf{C}$ and $\mathbf{E}^2$ in Equation 12. Assuming

that the population covariance of the residuals of the instruments is zero implies that $\mathbf{E}^2$ is a diagonal matrix with the variances of the residuals on the diagonal. From the assumption that the population covariances of all latent variables are zero, it follows that $\mathbf{C}$ is a diagonal matrix with the variances of the latent variables on the diagonal; assuming that the population covariances of only some of the latent variables are zero implies that the matrix $\mathbf{C}$ contains covariance parameters and zeros. In practical applications the investigator should state his or her assumptions regarding the population covariances of the latent variables.

These points can be illustrated by the multitrait–multimethod matrix. The model for this matrix is Equation 3: r_{jk} are the loadings on the trait factors and c_{jk} on the method factors. The matrix $\mathbf{L}$ contains a column for each trait factor and a column for each method factor. Each instrument has exactly one parameter for the corresponding trait factor and one parameter for the corresponding method factor (Jöreskog, 1971). In the most general formulation it is assumed that all factors are correlated (Werts & Linn, 1970) and that all elements of the matrix $\mathbf{C}$ are parameters. A special case is obtained by assuming that the trait factors are correlated and that the method factors are correlated but that the trait factors are not correlated with the method factors (Alwin, 1974). The matrix $\mathbf{C}$ contains parameters and zeros. A further restriction can be made by assuming that the effects of a given method are constant (Alwin, 1974). This implies that some parameters in the matrix $\mathbf{L}$ are constrained: The parameters in a column for a method factor are set equal to each other.

Equations 2, 4, 5, and 6 all contain a general factor, whereas Equations 3, 7, and 8 do not contain such a factor. Equation 3 has been discussed in the literature, as it is the basic model for the multitrait–multimethod matrix. In many cases the incorporation of a general factor seems to be important. First, the general factor can be of theoretical interest. For example, Eiting (Note 1) used musicality tests constructed in a facet design; he used a facet with musical abilities and a facet with musical materials such as melody and rhythm. The theoretical interest in this study was in the comparison of the models in Equations 3 and 4. Equation 3 implies that the correlation matrix of the musicality tests can be explained by material factors and separate factors for musical abilities, whereas Equation 4 implies that next to the material factors only one musical ability is involved. Other theoretical studies comparing models with and without a general factor can easily be conceived. Suppose tests are considered in a design with two facets of Guilford's (1967) structure-of-intellect model and that all the tests belong to the same element of the third facet, for example, memory tests for the elements of the Cartesian product of the content and product facet. A point of theoretical interest in this case is the comparison of models with and without a general memory factor. Second, the comparison of models with and without a general factor is important from the point of view of test construction. For items constructed in a facet design that fit a model with a general factor, it makes sense to combine items in a total score measuring the general aspect of the construct; items that do not fit a model with a general factor should only be combined in subtest scores. Regarding the general factor, it cannot be stated a priori that this factor should be correlated or uncorrelated with other factors; this depends on the research problem or can be investigated empirically. In test-construction, it is generally useful to postulate that the general factor is uncorrelated with the other factors. But in a theoretical study, such as in the above-mentioned structure-of-intelligence tests, it can be of interest to study the correlation of a general factor with other factors.

Estimation, Testing, and Identification

If in a sample an unbiased estimate $\hat{\mathbf{S}}$ of the covariance matrix is obtained, the parameters can be estimated using the generalized least squares or the maximum-likelihood method. By assuming that the observed scores on the instrument have a multivariate normal distribution for both methods, statistics can be computed that are asymptotically distributed as chi-squares (Jöreskog & Goldberger, 1972). The chi-square value can be used for testing the goodness of fit of a model against the very broad alternative that $\mathbf{S}$ is any positive, de-

Table 2
Means, Standard Deviations, and K-R 20 Reliability Coefficients for 5-Item Tests Measuring Social Anxiety in Children

Test	*M*	*SD*	K-R 20
Social–cognitive	.82	1.08	.52
Social–avoidance	1.23	.99	.11
Social–physiological	.65	.94	.49
Social–affective	.72	1.00	.45
Intellectual–cognitive	1.27	1.43	.65
Intellectual–avoidance	.74	.80	.11
Intellectual–physiological	1.23	1.39	.65
Intellectual–affective	1.17	1.12	.47
Physical–cognitive	1.04	1.27	.65
Physical–avoidance	.47	.82	.44
Physical–physiological	.54	.87	.47
Physical–affective	.42	.73	.42
Exclusion–cognitive	1.07	1.19	.56
Exclusion–avoidance	.62	.84	.31
Exclusion–physiological	.47	.93	.66
Exclusion–affective	1.71	1.16	.33
Appearance–cognitive	1.16	1.27	.56
Appearance–avoidance	1.37	1.00	.26
Appearance–physiological	1.24	1.32	.62
Appearance–affective	1.63	1.29	.46

Note. K-R 20 = Kuder-Richardson formula. For *M*s and *SD*s, $N = 320$; for K-R 20, $N = 396$: The group of 320 subjects was increased with the addition of 76 third-grade students.

definite matrix (Jöreskog, 1974). Moreover, different models can be compared by the right-tail probabilities of their chi-square values. Also, a Model M_1, nested in Model M_2, can be tested against Model M_2: The difference in chi-square values is asymptotically a chi-square with degrees of freedom equal to the difference of the degrees of freedom of the models. Model M_1 is said to be nested in M_2 if M_1 can be obtained from M_2 by constraining one or more of the free parameters in M_2 to be fixed or to be equal to one another (Long, 1976). Another method for the investigation of the fit of a model is the inspection of the matrix with the residual covariances:

$$\mathbf{R} = \hat{\mathbf{S}} - (\hat{\mathbf{L}}\hat{\mathbf{C}}\hat{\mathbf{L}}' + \hat{\mathbf{E}}^2), \qquad (13)$$

where $\hat{\mathbf{L}}$, $\hat{\mathbf{C}}$, and $\hat{\mathbf{E}}^2$ are the estimates for the matrices $\mathbf{L}$, $\mathbf{C}$, and $\mathbf{E}^2$. The computations can be done with the computer programs ACOVS (Jöreskog, Gruvaeus, & Van Thillo, Note 2) and LISREL (Jöreskog & Van Thillo, Note 3; Saris, Note 4).

A model is said to be identified if all parameters are uniquely determined. For example, the factor analytic model is not identified; the parameters are not unique, and the factor solution can be rotated. A necessary condition for identification is that the number of distinct parameters in the covariance matrix $\mathbf{S}$ be at least as high as the number of parameters implying that the number degrees of freedom must be equal to or greater than zero. This condition is necessary but not sufficient. A second condition is that the linear equations in the model be such that each individual parameter can be separated from the other parameters. In general it is hard to investigate this condition. Moreover, it is possible that the condition may be fulfilled but the model still not identified. This can happen if the estimate of a parameter in a sample is about zero (Saris, Note 4). A practical procedure for investigating the identifiability of a model is to ask for the standard errors of the parameter estimates. If the program can compute the standard errors, one has good assurance as to the identifiability of the model (Wiley, 1973). Another practical procedure is to estimate the parameters twice by starting the computing procedure with different initial values for the parameters. If the estimates of the parameters for both computer runs are equal, one also has good assurance as to the identifiability of the model (Saris, Note 4).

Measurement of Social Anxiety in Children

Dekking and Raadsheer (Note 5) used a situational and a behavioral facet for the construct of social anxiety in children. In the situational facet five types of situation in which children can exhibit social anxiety were differentiated: social, intellectual, physical, exclusion, and appearance. In the behavioral facet four types of reaction indicating social anxiety were differentiated: cognitive, avoidance, physiological, and affective. The Cartesian product of the two facets has 20 elements (see Table 1); for each element 5 items were written. An example of a social–avoidance item is as follows: "During recreation I don't play with other children" (yes/no); an example of an appearance–cognitive item is as follows: "Having had a hair cut I am afraid others think I look strange" (yes/no). The 100

items were administered to 320 children from the fourth, fifth, and sixth grades of three elementary schools. Each item was scored 1 (indicating social anxiety) or 0 (indicating the absence of social anxiety). For every 5-item test the score was the sum of the item scores. For the 20 tests, means, standard deviations, and Kuder-Richardson formula (K-R 20) reliability coefficients of the scores were calculated (see Table 2); the correlation coefficients among the 20 variables were also calculated (see Table 3). The interest in this study was in the relations among the variables and not in the scale in which the variables are measured. It was therefore decided to analyze the correlation matrix that is the covariance matrix of the standard scores. From the means in Table 2 it is obvious that the frequency distribution of most of the variables is skewed. The assumption of a multivariate normal distribution of the observed scores is certainly not fulfilled, and goodness of fit tests should therefore be interpreted very carefully. It is assumed that the population covariances of the residuals of the instruments are zero; therefore, in all the analyses $\mathbf{E}^2$ is a diagonal matrix with the residual variances on the diagonal.

The variance component model (Equation 9) is the model with the smallest number of parameters. Hence the analysis was started with this model; the form of Equation 11 is presented in Table 4. From the assumption that the population covariances of all latent variables are zero, it follows that in Equation 12 $\mathbf{C}$ is a diagonal matrix with diagonal elements: var (S), var (R_1), var (R_2), var (R_3), var (R_4), var (R_5), var (C_1), var (C_2), var (C_3), and var (C_4). Matrix $\mathbf{L}$ is not of full column rank; therefore, constraints are necessary for estimating the variance components (Jöreskog, 1974). In this stage of the investigation, however, interest was not in estimating the parameters but in fitting a model. The unconstrained model was therefore fitted to the data. The value of the chi-square for this model is 547.39 with 181 degrees of freedom; the fit of the variance component model is very poor.

In Equation 11 the variances of the latent variables of the vector $\mathbf{f}_i$ were subsequently set equal to one. The matrix $\mathbf{L}$ is then a matrix with loadings: The ones in the matrix $\mathbf{L}$ in Table 4 are replaced by parameters that should be estimated. In this case Equation 12 gives the structure of the covariance matrix of the observed scores in a restricted factor model

Table 3
Correlation Coefficients Among 20 Variables Measuring Social Anxiety in Children

Variable	y_{12}	y_{13}	y_{14}	y_{21}	y_{22}	y_{23}	y_{24}	y_{31}	y_{32}	y_{33}	y_{34}	y_{41}	y_{42}	y_{43}	y_{44}	y_{51}	y_{52}	y_{53}	y_{54}
y_{11}	.24	.45	.50	.57	.26	.44	.45	.34	.21	.43	.26	.56	.30	.39	.42	.52	.39	.53	.52
y_{12}		.13	.15	.19	.19	.19	.13	.19	.26	.15	.23	.19	.23	.14	.11	.19	.14	.14	.14
y_{13}			.54	.52	.24	.48	.47	.11	.11	.49	.15	.43	.25	.53	.34	.44	.31	.55	.43
y_{14}				.54	.28	.46	.54	.19	.09	.41	.25	.44	.41	.47	.38	.52	.39	.57	.53
y_{21}					.35	.64	.63	.24	.13	.35	.14	.57	.22	.39	.49	.59	.40	.56	.54
y_{22}						.31	.33	.14	.23	.17	.23	.20	.35	.20	.23	.25	.27	.27	.24
y_{23}							.59	.16	.17	.41	.16	.47	.28	.48	.46	.50	.33	.56	.49
y_{24}								.14	.11	.39	.13	.45	.28	.45	.43	.46	.34	.53	.48
y_{31}									.46	.30	.24	.28	.10	.13	.26	.27	.23	.22	.28
y_{32}										.24	.40	.18	.18	.14	.17	.18	.24	.17	.11
y_{33}											.33	.41	.27	.46	.31	.33	.24	.38	.29
y_{34}												.24	.18	.18	.16	.19	.17	.09	.16
y_{41}													.18	.47	.49	.53	.31	.54	.52
y_{42}														.31	.20	.22	.20	.29	.24
y_{43}															.36	.43	.32	.50	.35
y_{44}																.45	.34	.48	.48
y_{51}																	.32	.60	.62
y_{52}																		.42	.40
y_{53}																			.68
y_{54}																			

Note. N = 320.

Table 4
Form of Equation 11

$$\mathbf{y}_i \quad = \quad \mathbf{m} \quad + \quad \mathbf{L} \quad \mathbf{f}_i \quad + \quad \mathbf{e}_i$$

$$\begin{bmatrix} y_{i11} \\ y_{i12} \\ y_{i13} \\ y_{i14} \\ y_{i21} \\ y_{i22} \\ y_{i23} \\ y_{i24} \\ y_{i31} \\ y_{i32} \\ y_{i33} \\ y_{i34} \\ y_{i41} \\ y_{i42} \\ y_{i43} \\ y_{i44} \\ y_{i51} \\ y_{i52} \\ y_{i53} \\ y_{i54} \end{bmatrix} = \begin{bmatrix} m_{11} \\ m_{12} \\ m_{13} \\ m_{14} \\ m_{21} \\ m_{22} \\ m_{23} \\ m_{24} \\ m_{31} \\ m_{32} \\ m_{33} \\ m_{34} \\ m_{41} \\ m_{42} \\ m_{43} \\ m_{44} \\ m_{51} \\ m_{52} \\ m_{53} \\ m_{54} \end{bmatrix} + \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} S_i \\ R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \\ C_{i1} \\ C_{i2} \\ C_{i3} \\ C_{i4} \end{bmatrix} + \begin{bmatrix} E_{i11} \\ E_{i12} \\ E_{i13} \\ E_{i14} \\ E_{i21} \\ E_{i22} \\ E_{i23} \\ E_{i24} \\ E_{i31} \\ E_{i32} \\ E_{i33} \\ E_{i34} \\ E_{i41} \\ E_{i42} \\ E_{i43} \\ E_{i44} \\ E_{i51} \\ E_{i52} \\ E_{i53} \\ E_{i54} \end{bmatrix}$$

with one general factor, five situation factors, and four reaction factors. From the assumption that the factors are orthogonal, it follows that the matrix **C** in Equation 12 is an identity matrix. The value of the chi-square for this model is 210.49 with 130 degrees of freedom. Comparing the statistics of both models shows that the fit of the factor analytic model is better than the fit of the variance component model. Inspection of the matrix **R** shows that residual correlation coefficients higher than .10 or lower than −.10 are between the following variables: social–avoidance and physical–cognitive (.113), social–avoidance and physical–avoidance (.131), social–avoidance and physical–affective (.166), social–affective and exclusion–avoidance (.117), intellectual–avoidance and physical–affective (.137), and physical–affective and appearance–physiological (−.108). In five of the six cases, tests measuring avoidance reactions are involved. Moreover, from Table 2 it is clear that the tests measuring avoidance reactions generally show low K-R 20 reliability coefficients. Therefore, the five tests measuring avoidance reactions were eliminated from the correlation matrix, the avoidance-reaction factor was also eliminated, and the revised data were reanalyzed. The resulting chi-square value is 99.59 with 60 degrees of freedom. However, the standard errors of the parameters could not be computed, indicating that the model is not identified. Inspection of the parameter estimates showed that the loading of social–cognitive on the factor Social and the loading of exclusion–affective on the factor Exclusion are both near zero. To obtain an identified solution these two loadings were constrained to be equal to each other. In the computing procedure we encountered serious difficulties: In the iteration process the loading of social–affective on the factor Affective became larger and larger than one, while the residual variance became smaller and smaller than zero. It was therefore decided to stop the iteration process. A cause for difficulties in obtaining an identified model can be that all factors are specified to be uncorrelated. Therefore, it was assumed that the situation factors are correlated between one another and that the reaction factors are correlated between one another; the correlations between situation and reaction factors and between the situation and reaction factors on the one hand and the general factor on the other are assumed to be zero. The fit of this model is excellent: The value of the chi-square is 33.57 with 47 degrees of freedom; the range of the coefficients of **R** is from −.047 to .050. The program computes standard errors of the parameter estimates, indicating that the model

Table 5
Symmetrical 90% Confidence Intervals for Parameter Estimates

Variable	Factor loading									Variance residual
	General	Social	Intellectual	Physical	Exclusion	Appearance	Cognitive	Physiological	Affective	
Social–cognitive	.27, .80	.19, .52	0	0	0	0	−.02, .73	0	0	.39, .54
Social–physiological	−.22, .63	.25, .56	0	0	0	0	0	.45, .80	0	.32, .50
Social–affective	−.03, .70	.29, .61	0	0	0	0	0	0	.29, .79	.30, .50
Intellectual–cognitive	.21, 1.04	0	−.68, 1.07	0	0	0	.17, 1.07	0	0	.04, .35
Intellectual–physiological	.06, .86	0	−.21, .33	0	0	0	0	.32, .93	0	.32, .47
Intellectual–affective	−.06, .84	0	−.32, .50	0	0	0	0	0	.44, .96	.25, .44
Physical–cognitive	.39, .64	0	0	.15, .43	0	0	−.45, .36	0	0	.51, .79
Physical–physiological	−.07, .65	0	0	.40, .67	0	0	0	.29, .73	0	.24, .50
Physical–affective	.12, .41	0	0	.29, .55	0	0	0	0	−.12, .32	.62, .86
Exclusion–cognitive	.27, .90	0	0	0	.02, .46	0	.05, .85	0	0	.31, .47
Exclusion–physiological	−.23, .68	0	0	0	.04, .61	0	0	.47, .85	0	.29, .52
Exclusion–affective	.31, .78	0	0	0	−.05, .26	0	0	0	−.01, .71	.48, .67
Appearance–cognitive	.26, .87	0	0	0	0	.14, .40	.08, .81	0	0	.34, .48
Appearance–physiological	.15, .88	0	0	0	0	.26, .56	0	.22, .88	0	.20, .32
Appearance–affective	.41, .89	0	0	0	0	.24, .55	0	0	−.06, .77	.22, .36

Note. $N = 320$.

Table 6
Symmetrical 90% Confidence Intervals for Factor Correlation Coefficient Estimates

Factor	1	2	3	4	5	6	7	8	9
1. General	1								
2. Social	0	1							
3. Intellectual	0	−.92, 1.91	1						
4. Physical	0	.19, .73	−6.28, 4.27	1					
5. Exclusion	0	.22, .95	−12.13, 8.63	−.13, .80	1				
6. Appearance	0	.41, .91	−5.57, 4.09	−.82, .29	−.71, .96	1			
7. Cognitive	0	0	0	0	0	0	1		
8. Physiological	0	0	0	0	0	0	.72, 1.04	1	
9. Affective	0	0	0	0	0	0	.63, 1.07	.77, 1.02	1

Note. $N = 320$.

is identified. Table 5 gives the symmetrical 90% confidence intervals for the loading estimates, and Table 6 gives the confidence intervals for the factor correlation estimates. The confidence intervals for the correlations between the factor Intellectual and the other situation factors are inadmissibly large. The concept of the specific factor Intellectual is not very appropriate for the description of the correlation matrix. The use of a general factor, uncorrelated with all other factors, seems appropriate: 10 instruments have substantial loadings on this factor. Also, it seems appropriate to distinguish the factors Social, Physical, Exclusion, and Appearance. The factor Social is correlated with the other three factors: The intervals for the correlations do not contain the value zero; the intervals also do not contain the value one, indicating that it makes sense to distinguish the factor Social from the other three factors. The intervals of the intercorrelations of the factors Physical, Exclusion, and Appearance all contain the value zero. Moreover, almost all loadings on these four situation factors are substantial. There is, however, not much reason to distinguish the reaction factors: The intervals of the intercorrelations of these three factors all contain the value one.

Discussion

The analysis of the social anxiety data has shown that covariance structure analysis can be an important tool in test construction. To construct an instrument for measuring general social anxiety, one can select and combine tests with high loadings on the general factor. It is also possible to construct composite instruments that measure special aspects of social anxiety. For example, a composite of the tests physical–physiological, and physical–affective, and physical–cognitive would constitute an instrument that measures social anxiety in physical situations.

The question of why covariance structure analysis should be used for the analysis of correlation matrices from facet designs can be raised. Conventional factor analytic methods, especially three-mode factor analysis (Tucker, 1963), could also be used. Some points are mentioned briefly. First, conventional factor analytic methods are mostly used for exploring the structure of a correlation matrix, whereas covariance structure analysis is appropriate for testing hypotheses about matrices. Second, in covariance structure analysis the parameters can be estimated with the maximum-likelihood method. This implies that the properties of the estimates are known; the desirable properties of maximum-likelihood estimates are described by Mood and Graybill (1963, p. 185). Moreover, from the standard errors of the parameter estimates, confidence intervals can be derived. The properties of the estimates in conventional factor analytic methods are not known, and it is also not possible to specify confidence intervals. Third, it should be emphasized that it is assumed that the observed scores have a multivariate normal distribution. In applications this assumption is sometimes even not approximately fulfilled, and the effect of this is not known. Finally, it should also be emphasized that there are many problems in the use of covariance structure analysis. To men-

tion some, it is sometimes hard to formulate hypotheses and identifiable models, the computing time can be high, some problems, such as the power of the test of a model, are not resolved, and the variables must be quantitative and their relations linear.

The model with 10 factors proved to be an adequate explanation for the correlation coefficients between the 15 social anxiety variables. This does not mean, however, that it is the only model that fits the data. For the correlation matrix of the 20 variables, for example, the following model was used: five situation factors, four reaction factors, and two correlated second-order factors, one a general situation factor and the other a general reaction factor. Each test has a loading on one situation and on one reaction factor; each situation factor has a loading on the general situation factor, and each reaction factor has a loading on the general reaction factor. The model is a restricted, second-order factor analysis model (Jöreskog, 1974). The total number of parameters in this model is one less than in the model with the 10 orthogonal, first-order factors. The value of the chi-square for this model is 207.16 with 131 degrees of freedom; the fit of the model is about the same as the model with the 10 orthogonal, first-order factors. Inspection of the matrix **R** shows that the coefficients higher than .10 or lower than −.10 are between the following variables: physical–cognitive and social–avoidance (.116), physical–avoidance and social–avoidance (.138), physical–affective and social–avoidance (.136), physical–avoidance and social–affective (−.108), physical–affective and intellectual–affective (−.111), exclusion–affective and physical–cognitive (.112), appearance–avoidance and physical–cognitive (.109), physical–cognitive and appearance–affective (.105), and physical–affective and appearance–physiological (−.135). Although the picture is not as clear as in the previous analysis, it is likely that elimination of some tests can improve the fit of the model. This was not done because a model with two correlated second-order factors is more difficult to interpret than a model with first-order factors.

It has already been stated that the general model in Equation 2 can easily be generalized to more than two facets. For example, the model for the score of Subject i on the combination of the jth element of the first facet, the kth element of the second facet, and the lth element of the third facet is

$$y_{ijkl} = m_{jkl} + s_{jkl}S_i + r_{jkl}R_{ij} + c_{jkl}C_{ik} + h_{jkl}H_{il} + rc_{jkl}RC_{ijk} + rh_{jkl}RH_{ijl} + ch_{jkl}CH_{ikl} + E_{ijkl}. \quad (14)$$

From this equation it can be seen that the number of parameters is very large; restrictions in the model are necessary.

Reference Notes

1. Eiting, M. H. *De cognitie van muzikale transformaties: Een theoretische en multivariate benadering.* Unpublished manuscript, University of Amsterdam, Psychology Laboratory, Amsterdam, The Netherlands, 1977.
2. Jöreskog, K. G., Gruvaeus, G. T., & Van Thillo, M. *ACOVS—A general computer program for analysis of covariance structures* (Research Bulletin 70-15). Princeton, N.J.: Educational Testing Service, 1970.
3. Jöreskog, K. G., & Van Thillo, M. *LISREL—A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables* (Report No. 73-5). Uppsala, Sweden: University of Uppsala, Department of Statistics, 1973.
4. Saris, W. E. *Introduction to the use of linear structural equation models in non-experimental research and the LISREL-program.* Amsterdam, The Netherlands: Free University, 1977.
5. Dekking, Y. M., & Raadsheer, A. *Konstruktie van een instrument voor het meten van sociale angst bij kinderen.* Unpublished manuscript, University of Amsterdam, Psychology Laboratory, Amsterdam, The Netherlands, 1977.

References

Alwin, D. F. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner (Ed.), *Sociological methodology: 1973–1974.* San Francisco: Jossey-Bass, 1974.

Bentler, P. M. Multistructure statistical model applied to factor analysis. *Multivariate Behavioral Research,* 1976, *11,* 3–22.

Bock, R. D., & Bargmann, R. E. Analysis of covariance structures. *Psychometrika,* 1966, *31,* 507–543.

Butt, D. S., & Fiske, D. W. Comparison of strategies in developing scales for dominance. *Psychological Bulletin,* 1968, *70,* 505–519.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81–105.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral*

measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Endler, N. S., & Hunt, J. McV. Sources of behavioral variance as measured by the S-R inventory of anxiousness. *Psychological Bulletin*, 1966, *65*, 336–346.

Fiske, D. W. *Measuring the concepts of personality*. Chicago: Aldine, 1971.

Foa, U. G. New developments in facet design and analysis. *Psychological Review*, 1965, *72*, 262–274.

Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.

Guttman, L. Integration of test design and analysis. In, *Proceedings of the 1969 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1970.

Guttman, L., & Schlesinger, I. M. Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement*, 1967, *27*, 569–580.

Hamersma, R. J., Paige, J., & Jordan, J. E. Construction of a Guttman facet designed cross-cultural attitude-behavior scale toward racial-ethnic interaction. *Educational and Psychological Measurement*, 1973, *33*, 565–576.

Jöreskog, K. G. A general model for analysis of covariance structures. *Biometrika*, 1970, *57*, 239–251.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. *Psychometrika*, 1971, *36*, 109–133.

Jöreskog, K. G. A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Seminar Press, 1973.

Jöreskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 2. Measurement, psychophysics, and neural information processing*. San Francisco: Freeman, 1974.

Jöreskog, K. G., & Goldberger, A. S. Factor analysis by generalized least squares. *Psychometrika*, 1972, *37*, 243–260.

Lawley, D. N., & Maxwell, A. E. *Factor analysis as a statistical method* (2nd ed.). London: Butterworth, 1971.

Levin, J. Three-mode factor analysis. *Psychological Bulletin*, 1965, *64*, 442–452.

Long, J. S. Estimation and hypothesis testing in linear models containing measurement error. *Sociological Methods & Research*, 1976, *5*, 157–206.

Mood, A. M., & Graybill, F. A. *Introduction to the theory of statistics* (2nd ed.). New York: McGraw-Hill, 1963.

Mukherjee, B. N. Analysis of covariance structures and exploratory factor analysis. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 125–155.

Schlesinger, I. M., & Guttman, L. Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin*, 1969, *71*, 95–100.

Searle, S. R. *Linear models*. New York: Wiley, 1971.

Tucker, L. R. Implications of factor analysis of three-way matrices for measurement of change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

Werts, C. E., & Linn, R. L. Path analysis: Psychological examples. *Psychological Bulletin*, 1970, *74*, 193–212.

Wiley, D. E. The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Seminar Press, 1973.

Wiley, D. E., Schmidt, W. H., & Bramble, W. J. Studies of a class of covariance structure models. *Journal of the American Statistical Association*, 1973, *68*, 317–323.

Received March 13, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 777-803

Models for Biases in Judging Sensory Magnitude

E. C. Poulton
Medical Research Council, Applied Psychology Unit
Cambridge, England

Category ratings and magnitude judgments are affected by four range biases, the centering bias, the stimulus and response equalizing biases, and the contraction bias; by three nonlinear biases, the local contraction bias, the stimulus spacing bias, and the logarithmic bias; and by bias from transfer. Models of the biases are described. The biases are most marked in sensory dimensions that students are not taught to handle, such as loudness and brightness. Avoiding all the biases requires exceedingly rigorous investigations.

Why Sensory Magnitudes

Sensory magnitudes are selected for this review of biases in judgment because the stimuli can be measured on a physical scale. Judgments of the quality of life or of the likableness of people lack a precise measure of the stimulus. Thus the biases are more difficult to specify exactly. Anderson (1974), Helson (1964), Parducci (1963), and the late S. S. Stevens (1975) all use sensory magnitudes when presenting their theories of judgment.

Kinds of Stimuli

Averaging two weights and multiplying two lengths to calculate the area of a rectangle are a part of common knowledge. Students can handle these problems with relatively little bias (Anderson, 1974, Figures 1 and 8) because they are taught models of the way the sensory dimensions work.

Common knowledge does not include averaging two sound intensities or two light intensities, which is usually described as bisecting the interval between the two intensities. Nor does common knowledge include halving and doubling loudnesses and brightnesses. Judgments of these unfamiliar kinds are more easily biased. They are therefore the judgments to use in studying biases.

Kinds of Responses

Responses that are closely linked to the stimuli by well-known rules are less easy to bias than responses that are only loosely related. The display that follows lists the kinds of responses that are asked for in studying judgments of sensory magnitude. They are ordered by the extent to which they are linked closely to the stimuli by well-known rules:

Familiar physical units
Named categories
Numbered categories
Numbers
Cross-modal matches

First is a familiar physical measure of the stimulus such as length in meters. Provided the observer is given an adequate opportunity to study the stimulus, these responses are not easily biased except by the contraction bias. Next are rating scales with named or numbered categories. Finally, there are the numerical magnitude judgments and the cross-modal matching responses advocated by

The author is grateful to R. F. Fagot and R. Teghtsoonian for permission to make use of their unpublished information, to R. D. Patterson for commenting on a draft of the article, and to the British Medical Research Council for financial support.

Requests for reprints should be sent to E. C. Poulton, Medical Research Council, Applied Psychology Unit, 15 Chaucer Road, Cambridge, England CB2 2EF.

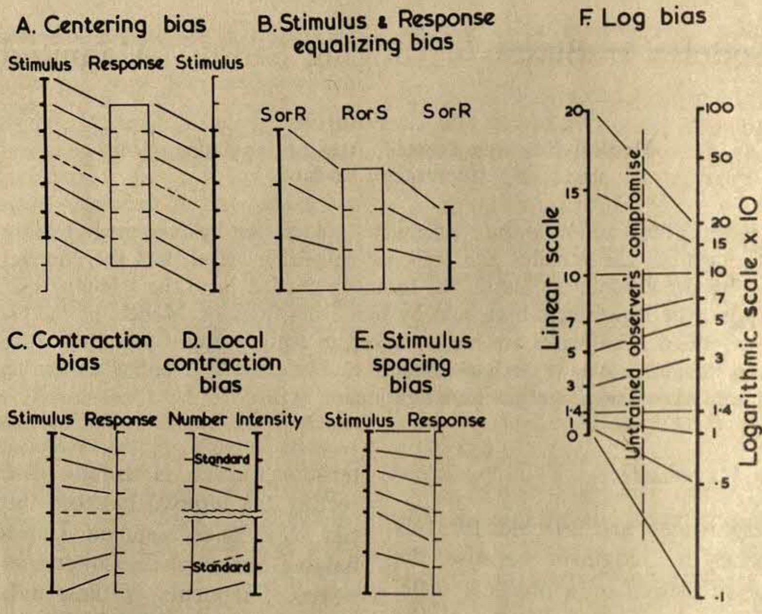

Figure 1. Models for the biases in judging sensory magnitudes; Models A, B, and C are concerned with the overall range, whereas Models D, E, and F are concerned with the nonlinearities within the overall range. (S = stimulus; R = response.)

S. S. Stevens (1975). Ratings and magnitude judgments are all relatively easily biased.

It is assumed in the display that the investigator selects the physical stimuli and that the observer gives one of the kinds of responses mentioned. But there is an alternative type of investigation in which the investigator gives one of the kinds of responses and the observer sets the size of the physical stimulus to match it. Both types of investigation can be biased in similar ways.

Kinds of Bias

In the centering bias of Figure 1A, the observer centers his range of responses on the range of stimuli. In the stimulus spacing bias of Figure 1E, the observer responds as if all the stimuli were equally spaced geometrically and equally probable. Parducci (1963; Parducci & Perrett, 1971) describes the effect of these two biases together by his range–frequency model. The model extends and modifies Helson's (1964) original adaptation-level model of the two biases.

The contraction bias of Figure 1C is a general characteristic of human behavior. It affects the distances reproduced by a limb as well as the judgments of anything that can be quantified, such as overt judgments of sensory magnitude (Poulton, 1973, 1974, 1975). Large stimuli and differences between stimuli are underestimated, whereas small stimuli and differences are overestimated. Once the observer knows the range of responses, he selects a response too close to the middle of the range. S. S. Stevens and Greenbaum (1966) call the contraction bias the "regression effect."

The local contraction bias of Figure 1D appears to have been described only for judgments of loudness. Acoustic stimuli in small, very high-intensity and very low-intensity ranges are treated as if they had less extreme values than they do have. This corresponds to the time–order error in investigations of the differential threshold (Hollingworth, 1910; Woodworth, 1938, pp. 438–448).

Judgments of sensory magnitude are affected also by the stimulus and response equalizing biases of Figure 1B. For the stimulus equalizing bias, the two scales at the sides represent stimuli, whereas the scale

in the middle represents responses. The observer uses his full range of responses whatever the size of the range of stimuli. For the response equalizing bias, the two scales at the sides represent responses, whereas the scale in the middle represents stimuli. The observer uses a larger range of responses when he is provided with a larger range.

Numerical judgments that require a step change in the number of digits are affected also by the logarithmic bias of Figure 1F. In using numbers logarithmically, the observer treats 1-, 2-, 3-, . . . , *n*-digit numbers as equally frequent instead of treating *n*-digit numbers as 10 times as frequent as (*n* − 1)-digit numbers. This shrinks the upper part of the numerical scale (Banks & Hill, 1974).

The logarithmic bias, the two contraction biases, and the stimulus and response equalizing biases can all affect the observer's very first judgment. In contrast, the effects of the centering bias and the stimulus spacing bias come on gradually as the observer learns the set of stimuli. These last two kinds of bias can both be said to be due to transfer from previous stimuli.

Judgments of sensory magnitude can be biased also by transfer from previous investigations, from instructions and demonstrations, from previous stimuli, from previous judgments, and from previous responses. The importance of the bias introduced into experiments by transfer is now becoming clear (Erlebacher, 1977; Greenwald, 1976; Poulton, 1973, 1974, 1975; Poulton & Freeman, 1966).

Centering Bias

The centering bias affects category ratings. Together with the stimulus spacing bias, it accounts for Helson's (1964) level of adaptation. Parducci (1963; Parducci & Perrett, 1971) shows that the level of adaptation depends on two characteristics of the distribution of stimuli, which can be varied independently. The influence of the midpoint of the range of stimuli is described here as the centering bias. The influence of the median of the distribution of stimuli is discussed later as the stimulus spacing bias.

Figure 1A illustrates a model of the centering bias. The stimulus intensities used in the theoretical investigation represented on the left side of the figure are two steps greater than the stimulus intensities used in the investigation represented on the right. Yet once the observer has learned the range, he centers his response scale on the range of stimulus intensities, whatever their physical values are. This is indicated by the lines that all slope down toward the right. Thus, a more intense stimulus from the investigation represented on the left side of the figure receives the same rating as a less intense stimulus from the investigation represented on the right side of the figure.

Centering Bias in Rating Noisiness

Figure 2 illustrates the centering bias in five fairly comparable investigations of the noises made by road vehicles and by aircraft. The filled points represent motor vehicle noise (from D. W. Robinson, Copeland, & Rennie, 1961). The unfilled points represent aircraft noise (from Bowsher, Johnson, & Robinson, 1966). The vertical lines show the range of noises in decibels (dB[A]). The short horizontal lines show the midpoints of the ranges. The vertical lines have been separated horizontally to make all the midpoints of the ranges lie on the dashed straight line.

The diamonds indicate the midpoints of the straight lines fitted to the ratings. This is the transition between acceptable and noisy in the D. W. Robinson et al. (1961) investigation and between moderate and noisy in the two Bowsher et al. (1966) investigations. In the Andrews and Finch (1951) and Lauber (Note 1) investigations the diamonds indicate the transition between unobjectionable or inoffensive and objectionable or annoying.

If the observers were to behave like sound level meters and have a fixed criterion for the just acceptable level of noise, all the diamonds in Figure 2 would lie on a horizontal line. If the ratings were determined entirely by the midpoints of the ranges of sounds presented, the diamonds would be superimposed upon the short horizontal lines.

Table 1
Avoiding Biases

Bias	Avoiding bias in category rating	Avoiding bias in magnitude estimation
	Range biases	
Centering (Figure 1A) The observer centers his range of responses on the range of stimuli	Use only the set of responses that has a demonstrably unbiased center, or Ask for only a single judgment	Ask for only a single judgment
Stimulus equalizing (Figure 1B) Whatever the size of the stimulus range, the observer uses his full range of responses	Unavoidable	Use a stimulus dimension of the same subjective size as the response dimension
Response equalizing (Figure 1B) Whatever size of response range the observer is given, he distributes the responses over the stimulus range	Unavoidable	Use a response dimension of the same subjective size as the stimulus dimension
Contraction (Figure 1C) The observer underestimates large stimuli and differences between stimuli and overestimates small stimuli and differences	Use only the neutral rating of the stimulus at the center of the range of named categories, or	Balance the biases by interchanging stimuli and responses and taking the average
Once the observer knows the range of responses, he selects a response too close to the middle of the range	Anchor the rating scale to the stimulus range at both ends	
	Nonlinear biases	
Local contraction (Figure 1D) The observer treats stimuli in small very high-intensity and very low-intensity ranges as if they had less extreme values		Avoid a small stimulus range with extreme values
Stimulus spacing (Figure 1E) The observer responds as if the stimuli were equally spaced geometrically and equally probable	Use equal geometric spacing and present the stimuli equally often, or Ask for only a single judgment	Use equal geometric spacing and present the stimuli equally often, or Ask for only a single judgment
Logarithmic (Figure 1F) The observer treats 1-, 2-, 3-, . . ., n-digit numbers as if they were equally frequent; this shrinks the upper part of the numerical scale	Stick to one-digit numbers	Stick to a range of numbers all having the same number of digits

Table 1 (*continued*)

Bias	Avoiding bias in category rating	Avoiding bias in magnitude estimation
	Transfer biases	
Transfer from previous investigations Transfer from instructions and demonstrations Transfer from previous stimuli Transfer from previous judgments Transfer from previous responses	Ask unpracticed observers for only a single judgment with unbiased instructions and no demonstrations	Ask unpracticed observers for only a single judgment with unbiased instructions and no demonstrations
	All biases combined	
	Ask separate groups of unpracticed observers for only a single judgment with unbiased instructions and no demonstrations; then use only the neutral rating of the stimulus at the center of the range of named categories; the stimulus and response equalizing biases are unavoidable	Ask unpracticed observers for only a single judgment with unbiased instructions and no demonstrations; also avoid a small stimulus range with extreme values and a response range with a step change in the number of digits; also use for stimuli and responses two sensory dimensions with subjectively equal-sized ranges, and balance the contraction biases by interchanging stimuli and responses for a separate group of unpracticed observers and averaging the results

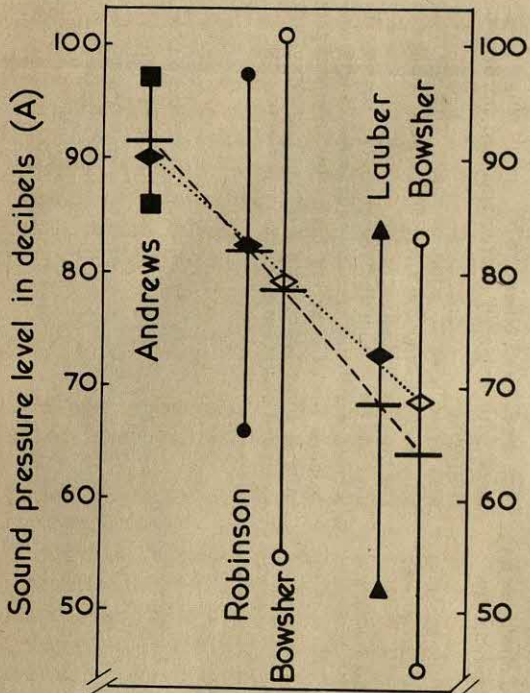

Figure 2. The centering bias in rating noises. (The vertical lines indicate the ranges of noise intensities heard. The short horizontal lines represent the midpoints of the ranges. The diamonds show the average just acceptable noise levels that lie at the midpoints of the rating scales obtained from separate groups of between 19 and 37 observers. Andrews = Andrews & Finch, 1951; Robinson = D. W. Robinson, Copeland, & Rennie, 1961; Bowsher = Bowsher, Johnson, & Robinson, 1966; and Lauber = Lauber (Note 1). Adapted from Poulton, 1977, Figure 2.)

The dotted line fitted to the diamonds by eye is a compromise, which favors the midpoints of the ranges. The midpoints of the ratings are determined less by the actual sound levels than by the ranges.

In the Andrews and Finch investigation on the left side of Figure 2, the diamond representing the midpoint of the 10-point scale of objectionableness is too high. It is pulled up by the middle of the range of noise intensities. In the investigations illustrated on the right side of the figure, the diamonds are too low. They are pulled down by the middle of the range of noise intensities. The dashed and dotted lines cross at about 85 dB (A). Here a diamond would be pulled neither up nor down. The crossover represents the only point in the data that is not affected by the centering bias.

Avoiding the Centering Bias in Ratings

A field investigator who wants to obtain named categories to represent sensory magnitudes can avoid the centering bias by using a method that provides data comparable to all the data illustrated in Figure 2. Separate groups of observers hear all the sounds, each group with a different constant amplification or attenuation. Thus, each group of observers provides a single vertical function like one of those of Figure 2. For the group whose average midrating is found to lie at the crossover point illustrated in the figure, the judgments are not affected by the centering bias. This method of avoiding the centering bias is listed at the top of the middle column of Table 1.

Table 1 indicates that the centering bias can also be avoided by restricting each group of observers to a single stimulus. If the investigator wants to measure only the just acceptable noise level at the middle of the range of ratings, separate groups of observers can each be given a single noise to judge at a different intensity. The observers have to decide simply whether the noise is acceptable or not acceptable. The noise that is judged acceptable by just 50% of the observers can then be determined. However, it is essential that the observers not have heard a range of noises of different intensities before judging their single noise. This is because their judgment will be affected by the previous stimulus range, even though they do not judge the noises (Bowsher et al., 1966, Table 3; Von Wright & Kekkinen, 1970).

Investigators who use numerical category ratings are often concerned only with the relative positions of the sensory magnitudes on the rating scale, not with their absolute positions. Here the centering bias does not matter. Often the investigator anchors the upper end of the rating scale to the most intense stimulus in his series and the lower end of the rating scale to his least intense stimulus. In doing so, he centers the rating scale on the range of stimuli.

Centering Bias in Direct Magnitude Estimation

The top of the last column of Table 1 indicates that in direct magnitude estimation, the centering bias can also be avoided by asking each observer for a single judgment. However, in using direct magnitude estimation, the investigator generally wishes to obtain only a relative judgment between two intensities, not an absolute judgment. The relative judgment shows how the reported difference in sensation varies with the difference in intensity. S. S. Stevens (1971, 1975) uses the following model:

$$\psi = K\,\phi^n,$$

or taking logs,

$$\log \psi = \log K + n \log \phi,$$

where ψ is the subjective intensity and ϕ is the physical intensity. In this model the centering bias is represented by log K. It is equated across observers or investigations by fixing or adjusting the value of K.

Logarithmic Bias

The Arabic numbering system introduces a logarithmic bias into the observer's use of numbers when he reaches a step change in the number of digits. After counting up to 10 the observer has two alternatives for the way in which to proceed. He can behave linearly and continue 11, 12, 13, . . . , or he can be more logarithmic and continue 20, 30, 40, ...

If the observer generates numbers linearly, he uses two-digit numbers 10 times as often as single-digit numbers. He uses three-digit numbers 100 times as often, and so on. If the observer generates numbers logarithmically, he uses 1-, 2-, 3-, . . . , n-digit numbers equally often. This shrinks the upper part of the response scale, as illustrated on the right side of Figure 1F.

The numbers generated can be arranged in rank order of size at equal intervals along the abscissa of a graph. When the numbers are plotted with a linear scale on the ordinate, the function is markedly concave upward, whereas when the scale on the ordinate is logarithmic, the function is slightly concave downward. The untrained observer compromises between pure linear and pure logarithmic choices, but favors logarithmic choices (Banks & Hill, 1974).

Figure 1F shows the relationship between numbers plotted logarithmically and numbers plotted linearly. The two sets of numbers are scaled so that they are the same height at 10 and at 1.4 on the two scales. At these two points,

$$x = 10 \log_{10} x.$$

At other points, the corresponding numbers are not the same height because here x and $10 \log_{10} x$ are different.

The lines connect various xs on the linear scale to the corresponding $10 \log_{10} x$s on the logarithmic scale. Between 3 and 7 the lines are almost parallel, so the two scales are virtually identical. Above 7 the logarithmic scale shrinks compared with the linear scale, while below 3 it expands compared with the linear scale. But when observers generate numbers in an experiment, the variability is so large that it is usually not possible to distinguish reliably between the two numbering systems in the range between 1 and 10.

The figure shows that above 10 the logarithmic scale is very much more condensed than the linear scale. So when an observer generates numbers that extend from 1 or 2 up to 50 or 100, it is possible to determine whether he is behaving linearly or logarithmically. The broken vertical line in Figure 1F indicates that in generating numbers, the unpracticed observer compromises between the two scales.

Numbers below 1.0 are rarely used in psychophysical investigations. Here the logarithmic scale is very much more spread out than the linear scale. This explains why when sensory magnitude is plotted with a logarithmic scale on the ordinate of a graph against physical magnitude on the abscissa, the slope of any psychophysical function becomes steeper as the threshold is approached. The threshold represents zero sensory magnitude. On the logarithmic scale of the ordinate, the threshold is infinitely far away in a downward direction (Poulton, 1968, p. 5).

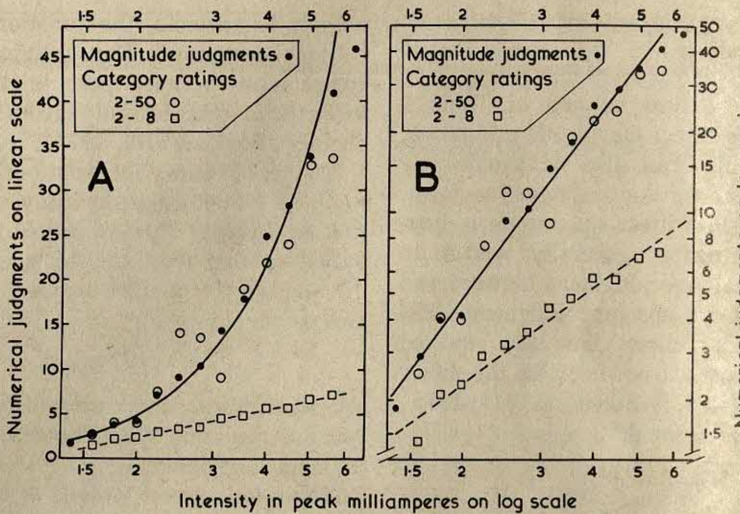

Figure 3. The logarithmic bias in direct magnitude estimation of electrotactile stimuli and in rating the stimuli on a 49-point scale corresponding to the average range of numbers used in the magnitude estimation. (Ratings using 7 categories are also shown. Panel A has a linear scale on the ordinate, whereas Panel B has a logarithmic scale. Results are from Gibson & Tomko, 1972.)

Logarithmic Bias in Rating

Figures 3A and 3B (results of study by Gibson & Tomko, 1972) illustrate the logarithmic bias that occurs when an observer rates electrotactile intensity with a large number of numerical categories. The numerical judgments on the ordinate are plotted in Figure 3A with a linear scale and in Figure 3B with a logarithmic scale. Thus, the solid straight line of Figure 3B becomes the solid line that is concave upward in Figure 3A. The electrotactile stimuli on the abscissa have the same logarithmic scale in both figures.

The unfilled circles represent the arithmetic means of a group of 11 observers who use a rating scale with 49 main categories. The categories 1 and 51 are reserved for stimuli judged to fall outside the initially defined range of stimuli. With the linear vertical scale of Figure 3A, the function represented by the unfilled circles is markedly concave upward, whereas with the logarithmic vertical scale of Figure 3B, the function is slightly concave downward. The slight downward concavity is indicated by the unfilled circles at the top and bottom of the function, which lie below the solid straight

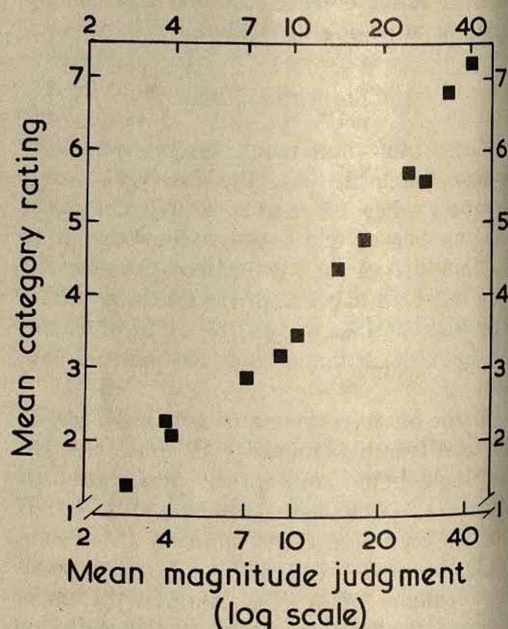

Figure 4. Mean 7-point category ratings plotted on a linear scale against mean magnitude judgments plotted on a logarithmic scale. (The two sets of data are from two separate groups of 11 observers. The function is slightly concave upward because the magnitude judgments are plotted on the appropriate logarithmic scale instead of on a linear scale. Results are from Gibson & Tomko, 1972.)

line. The slight downward concavity shows that the observers do not use a purely logarithmic numbering system, although their hybrid numbering system is nearly logarithmic.

The logarithmic bias does not affect ratings with 9 or less numerical categories (Montgomery, 1975; Parducci, 1963). The squares and the dashed fitted straight lines in Figures 3A and 3B represent the arithmetic means of a separate group of 11 observers who use a rating scale with 7 main categories. The categories 1 and 9 are reserved for stimuli judged to fall outside the initially defined range of stimuli. With the linear vertical scale of Figure 3A, the function represented by the squares is closely fitted by the straight line. Thus the logarithmic bias affects the rating scale with 49 categories but not the rating scale with 7 categories.

S. S. Stevens (1975, Figure 50) reports a logarithmic bias in ratings of loudness using seven categories. But this is presumably caused by transfer of the logarithmic bias from previous investigations on magnitude estimation performed by the same observers. Transfer between S. S. Stevens's various investigations is discussed later in this review.

With the logarithmic vertical scale of Figure 3B, the squares are also fitted fairly well by the straight line. This is because Figure 1F shows that within the range between 1 and 10, the linear and logarithmic scales are reasonably comparable. The most deviant square in Figure 3B is on the extreme left, with a mean rating as low as 1.4. Here Figure 1F shows that the two scales begin to differ appreciably.

Logarithmic Bias in Direct Magnitude Estimation

The filled circles and fitted solid line in Figure 3B illustrate the logarithmic bias in direct magnitude estimation. Here the points represent geometric means instead of arithmetic means. The results are from another separate group of 11 observers who judge the stimulus intensities in numbers without a specified standard or modulus. In Figure 3B the function represented by the filled circles is slightly concave downward, like the function represented by the unfilled circles for the ratings. The slight downward concavity shows that the observers do not use a purely logarithmic numbering system. However the filled circles in Figure 3A show that the observers certainly do not use a linear numbering system.

When the magnitude judgments of Figures 3A and 3B are plotted against the category ratings on the 49-point scale represented by the unfilled circles, the resulting function is a straight line (Gibson & Tomko, 1972, Figure 4). This indicates that the logarithmic bias affects magnitude judgments and category ratings to a similar extent, provided the range of numbers used as responses is about the same. In Gibson and Tomko's experiment, the numbers used for the category ratings are selected to correspond to the average range of numbers used in the magnitude judgments. It follows that the difference between direct magnitude estimation by unpracticed observers and numerical category rating is simply in the range of numbers normally used, not in the kind of judgment that the observers are instructed to make. Montgomery (1975, Figure 1, Panels A_1, A_2, and A_3) makes a similar point.

Ratings and Magnitude Estimates

The same stimuli can be judged while using either a linear numbering system or a logarithmic numbering system. When the two sets of judgments are plotted against each other with the linear scale on the ordinate and the logarithmic scale on the abscissa, the resulting function should be a straight line. But as Torgerson (1960) points out, the function is usually slightly concave upward, like the function in Figure 4. The figure shows Gibson and Tomko's (1972) 7-point ratings plotted on a linear scale against their magnitude judgments plotted on a logarithmic scale. The function is slightly concave upward because Figures 3A and 3B show that the ratings use a linear numbering system, whereas the magnitude judgments have a logarithmic bias, but are not completely logarithmic.

Torgerson's (1960, Figures 3 and 4) re-

sults for Munsell grays and Eisler's (1962, Figure 4, Panels ILa, IILa, IIILa, ISa, IISa, and IIISa) and Montgomery's (1975, Table 2) results for loudness all give functions that are slightly concave upward, like the function in Figure 4. Where the stimulus spacing does not bias the shape of the functions, Montgomery's results are clearly due to the hybrid numbering system used for the magnitude judgments, as with the data of Figure 4. Unfortunately Torgerson's and Eisler's results are influenced by the stimulus spacing bias and perhaps also by transfer from previous magnitude judgments or category ratings. Transfer bias from previous conditions is not possible in Gibson and Tomko's and Montgomery's experiments because they use a separate group of observers for each condition.

Figure 3A shows that Gibson and Tomko's (1972) 7-point ratings are linear on a lognormal or semilog plot. Thus in physical units, the equal intervals between the ratings on the ordinate of Figure 4 correspond to a logarithmic physical scale. The logarithmic scale of the magnitude judgments on the abscissa of Figure 4 also corresponds more or less to a logarithmic physical scale. This is because Figure 3B shows that the relationship between the magnitude judgments and the physical units is almost a straight line on a loglog plot. So both the vertical and the horizontal scales of Figure 4 are more or less equivalent to the logarithm of the physical units, hence the more or less straight-line relationship between them in the figure.

It is important to distinguish between the slightly concave upward relationship of Figure 4 and the markedly concave downward relationship reported by S. S. Stevens and Galanter (1957, Figures 8 and 9) between category and direct magnitude scales of loudness. S. S. Stevens and Galanter plot the logarithmic ratio judgments using a linear scale of sones instead of a logarithmic scale. This converts the slightly concave upward relationship of Figure 4 into S. S. Stevens's well publicized concave downward relationship (Eisler, 1962, Figure 4, Panels ILb, IILb, IIILb, ISb, IISb, and IIISb; Gibson & Tomko, 1972, Figure 3; Torgerson, 1960, Figure 3.1). The concave downward relationship is produced simply by plotting what corresponds to a logarithmic physical scale on the ordinate against what corresponds to a linear physical scale on the abscissa. It has no greater significance than this.

Avoiding Hybrid Numbering Systems

The logarithmic bias is most likely to occur when the observer uses a range of numbers that includes a step change in the number of digits. In S. S. Stevens's method of direct magnitude estimation, a stimulus in the middle of the range of stimuli may be called 10. The observer has then to use both one-digit and two-digit numbers. If the observer serves in several such investigations, the logarithmic bias in using numbers greater than 10 is likely to transfer to the use of numbers between 1 and 10. Eventually the observer may consistently use numbers logarithmically. Differences between sensory intensities are then expressed as numerical ratios. The consistent use of a purely logarithmic numbering system can be encouraged by instructing the observer to use ratios, as S. S. Stevens (1971, p. 428) does. Those who do not give ratios consistently are not invited to serve as observers.

The logarithmic bias should not occur when the observer uses numbers that all have an equal number of digits, except as a result of transfer from prior ratio judgments. The methods of avoiding the logarithmic bias are listed at the bottom of the second section of Table 1.

Stimulus Spacing Bias

The stimulus spacing bias is one of the two biases that together account for Helson's (1964) level of adaptation (Parducci, 1963; Parducci & Perrett, 1971), the other being the centering bias, which has already been discussed. Figure 1E illustrates a model of the stimulus spacing bias. The five stimuli at the top of the range of intensities on the left are spaced at geometric intervals only half the size of the three stimuli at the bottom of the range. In judging the stimuli,

the observer uses their rank order of magnitude rather than their relative subjective sizes. He behaves as if all the stimuli were subjectively equally spaced. In category rating this means using all the categories equally often. Thus the smaller intervals are overestimated compared with the larger intervals.

The *stimulus frequency bias* is a special case of the stimulus spacing bias. The observer behaves as if all the stimuli were equally probable. When one stimulus is presented more frequently than the remainder, the observer treats the more frequent stimuli as if they were all nearly, but not exactly, the same size. Thus, three identical stimuli one quarter of the way from the top of the stimulus range are treated as three closely spaced stimuli of slightly different size, as illustrated in Figure 1E. In judging all the stimuli while using their rank order of magnitude rather than their relative subjective sizes, the observer allocates an excessive amount of his response scale to the identical stimuli. The stimulus region around the identical stimuli is therefore overestimated compared with the remainder of the stimulus range.

Stimulus Frequency Bias in Rating

The stimulus frequency bias in category rating is described at length by Parducci (1963; Parducci & Perrett, 1971), who uses separate groups of observers for each condition and by Pollack (1964a, 1965a, 1965b), who uses the same or different observers. A number of extra stimuli of one particular intensity are added to a set of stimuli. The observer tends to use all his response categories equally often. Thus stimuli a little greater than the added stimuli receive higher ratings than previously, whereas stimuli a little smaller than the added stimuli receive lower ratings than previously. In the rating scale the added stimuli produce a local expansion, which is flanked on both sides by a compensatory contraction.

Stimulus Spacing Bias in Rating

An example of the stimulus spacing bias that corresponds to the model of Figure 1E

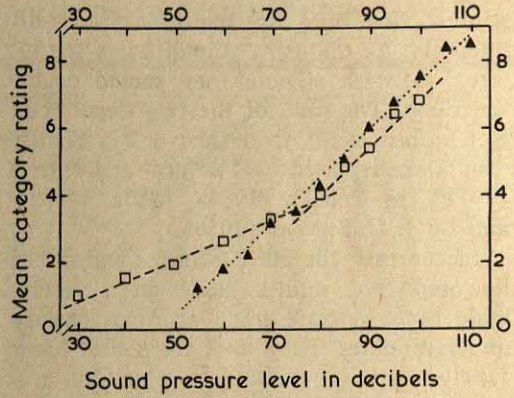

Figure 5. The stimulus spacing bias in rating the loudness of white noise. (Results are from Montgomery, 1975.)

is illustrated on the lognormal or semilog plot of Figure 5 (results from Montgomery, 1975). The unfilled squares represent ratings of the loudness of white noise made on a 7-point scale by 10 psychology undergraduates. The stimuli are spaced at 10-dB intervals between 30 and 100 dB (SPL), with additional stimuli inserted at 85 and 95 dB. The two inserted stimuli increase the slope of the function at the top end. The six points at 10-dB intervals are fitted by a separate

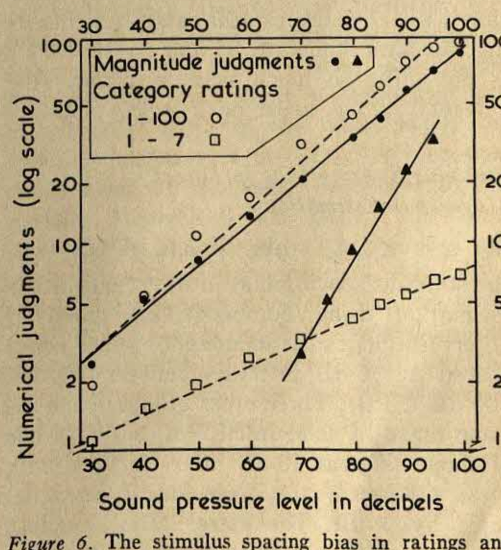

Figure 6. The stimulus spacing bias in ratings and direct magnitude judgments of the loudness of white noise. (The difference in slope between the filled triangles and the filled circles is produced by the stimulus equalizing bias. Results are from Montgomery, 1975.)

straight line from the five points at 5-dB intervals. As the extra stimuli halve the interval between stimuli, they should double the slope. The ratio of the two slopes is in fact rather larger than this: 1:2.3. Similar results are reported by Parducci and Perrett (1971), by Pollack (1964a, 1965a, 1965b), and by J. C. Stevens (1958).

In contrast, the filled triangles and dotted line represent ratings made on a 9-point scale by a separate group of 10 psychology undergraduates. They show the almost linear function that is produced by a uniform geometric spacing between all stimuli.

The unfilled squares and dashed fitted line of the loglog plot of Figure 6 (results from Montgomery, 1975) correspond to the unfilled squares and dashed fitted lines in Figure 5. Figure 6 shows that the discontinuity in the function of Figure 5 disappears when the ratings are plotted on a logarithmic scale. Inserting the extra stimuli between the stimuli at the top of the range helps to make the category ratings linear on a loglog plot like that of Figure 6.

The unfilled circles and dashed fitted line in Figure 6 are for category ratings between 1 and 100 made by another group of 10 psychology undergraduates. They show that with as many as 100 categories, inserting the extra stimuli at 85 and 95 dB (SPL) is not quite so much help in making the category ratings linear on a loglog plot.

Stimulus Spacing Bias in Direct Magnitude Estimation

J. C. Stevens (1958, Figure 3) describes the stimulus spacing bias in direct magnitude estimation using numbers. One group of observers judges the loudness of white noises spaced at 10-dB intervals between 40 and 100 dB (SPL). Three other groups judge the same range of noise intensities, but with one of three different 10-dB intervals filled with three or four extra noise intensities separated by 2 dB. The extra stimuli produce local increases in the steepness of the psychophysical function relating loudness to intensity on a loglog plot like that of Figure 6.

The stimulus spacing bias can be used to increase the effect of the logarithmic bias illustrated in Figure 1F. In judging sensory magnitudes, unpracticed observers produce functions on a loglog plot that are slightly concave downward, like the function represented by the filled triangles in Figure 6 (results from Montgomery, 1975) and by the filled circles in Figure 3B (Gibson & Tomko, 1972). The filled circles in Figure 6 are from a separate group of 10 psychology undergraduates who judge noises ranging from 30 to 100 dB (SPL). Here the stimuli are spaced 10 dB apart, except for additional stimuli inserted at 85 and 95 dB. The two inserted stimuli straighten out the concavity that would otherwise occur at the top end of the function.

Inserting additional stimuli near the top of the range is now a well-known procedure for obtaining a nearly straight-line function on a loglog plot with unpracticed observers (Eisler, 1962; J. C. Stevens & Tulving, 1957). This does not prove that subjective sensory intensity is a power function of physical intensity. It simply demonstrates the investigator's practical familiarity with the logarithmic and stimulus spacing biases and his skill in combining them to produce a nearly straight line on a loglog plot.

Avoiding the Stimulus Spacing Bias

The second section of Table 1 indicates that the stimulus spacing bias can be avoided by restricting each group of observers to a single judgment. A more practical alternative is to space all the stimuli geometrically and to present the stimuli equally often.

Confusions between stimuli reduce the slope of the psychophysical function at the point where the confusions occur. At the limit, where two different stimuli are always confused, the slope is zero because both stimuli receive the same average response. Local variations in slope can therefore be removed by making successive pairs of stimuli equally confusable. Following Weber's law (see Woodworth, 1938, pp. 430–438), this means using geometric spacing of stimuli. Geometric spacing is thus the subjectively neutral spacing, which produces a straight line and avoids the stimulus spacing bias.

Geometric stimulus spacing can be used in both category rating and in magnitude estimation. In category rating with not more than about nine categories, it produces approximately equally spaced ratings. In magnitude estimation by observers who are trained to use numbers logarithmically, it produces approximately equal subjective ratios, which are equally spaced on a logarithmic plot.

Contraction Biases

Figure 1C illustrates a model for the contraction bias. Large stimuli and differences between stimuli are underestimated, while small stimuli and differences between stimuli are overestimated. The contraction bias affects an observer's very first judgment, because he always has some idea of the range of stimuli or stimulus differences to expect.

The contraction bias is facilitated by giving the observer a limited set of responses to use with an obvious middle value. Examples are a set of ratings, the finite set of numbers used in fractional magnitude judgments, and the range of the gain control used in magnitude production and cross-modality matching. Once the observer knows the range of responses, he selects a response that is closer to the middle of the range of responses than it should be (S. S. Stevens & Poulton, 1956, Table 1).

Contraction Bias in Judging Visual Height

To investigate the contraction bias directly, it is necessary to obtain responses in the physical units used to measure the stimulus. Figure 7 illustrates the contraction bias in very first judgments of the height of a white post in a field (from Joynson, Newson, & May, 1965). The abscissa and the solid lines show the true height in inches. The points represent the mean very first judgments of five separate groups, each with 30 observers, in Experiment 1 and of four separate groups, each with 18 observers, in Experiment 2. On the left of the figure, the heights of short posts are overestimated. On the right, the heights of tall posts are underestimated. As in Figure 2, the only unbiased points are where the lines cross.

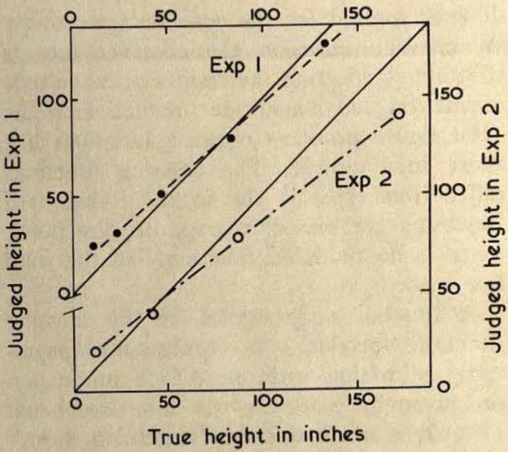

Figure 7. The contraction bias in the very first judgments of the height of a white post in a field. (One third of the observers in each group judge their post at a distance of 25 m, one third at 100 m, and the remaining third at 400 m in Experiment 1 and at 230 m in Experiment 2. Results are from Joynson, Newson, & May, 1965.)

Experiment 1 is carried out on an airfield. The dashed line fitted to the filled points crosses the solid line at a height of 80 inches (2 m). This is about the height of the perimeter fence around most airfields. Experiment 2 is carried out on a level field of mown grass. The fitted dashed line crosses the solid line at a height of 48 inches (1.2 m). This is close to the mean height of about 54 inches (1.4 m; SD = 12 inches or .3 m) given by 40 people in answer to the question, "How tall is a post?" (Joynson, Note 2). In making their very first judgments, the observers appear to take as a reference the expected height of a fence or post. They select a height a little nearer to this reference height than it should be. The reference height takes the place of the middle of the range in Figure 1C.

Contraction Biases in Direct Magnitude Estimation and Production

The contraction bias can be demonstrated indirectly by contrasting magnitude estimation with magnitude production. The difference between the two procedures is in the

dimension used for the observer's responses. Whichever dimension the observer uses is contracted. Plotting the results of magnitude estimation and magnitude production in the same figure produces crossing functions like those in Figure 7. The crossing functions differ from those in the figure in that both functions are biased in opposite directions; there is no unbiased function like the solid line in the figure.

A number of examples of the crossing functions produced by contrasting magnitude estimation with magnitude production are given by S. S. Stevens and Greenbaum (1966), who call this the "regression effect." However, in most of these investigations the two dimensions differ in size. Thus, the crossing functions are influenced by the stimulus and response equalizing biases as well as by the contraction bias. There are also transfer effects, because each observer makes a number of both kinds of judgment, each kind in a separate part of the experiment.

Sequential Contraction Bias

After the presentation of a stimulus near the top of the range of stimulus intensities, the next stimulus tends to be judged larger than it should be. After the presentation of a stimulus near the bottom of the range of intensities, the next stimulus tends to be judged smaller than it should be. This result is found in the rating of stimuli on a 10-category rating scale (Holland & Lockhead, 1968; Ward, 1972; Ward & Lockhead, 1971), in direct magnitude estimation without a standard (Cross, 1973; Jesteadt, Luce, & Green, 1977, Experiments 1 and 3), in direct magnitude estimation with a variable standard (Jesteadt et al., 1977, Experiment 2; Ward, 1973), and in cross-modal matches without a standard (Ward, 1975).

The sequential contraction bias is a special case of the contraction bias. After the presentation of a stimulus at one end of the range of stimuli, the average difference from the next stimulus selected at random is likely to be large. Large differences tend to be underestimated. Thus, the next stimulus will be judged on the average nearer than it should be to the previous stimulus. After the presentation of a stimulus near the top of the range, the bias produces a positive average constant error. After the presentation of a stimulus near the bottom of the range, the bias produces a negative average constant error. After the presentation of stimuli near the middle of the range, the average constant error is small.

When the subjective judgments are plotted against the physical stimuli, the sequential contraction bias reduces the slope of the psychophysical function. This is because after the presentation of a stimulus near the top end of the range, the stimulus with a large difference that is likely to be underestimated will be near the bottom end of the range. The size of the range is therefore underestimated. There is a similar underestimation after the presentation of a stimulus near the bottom end of the range, because the stimulus with a large difference that is likely to be underestimated will be near the top end of the range. Thus the sequential contraction bias provides the mechanism for S. S. Stevens and Greenbaum's (1966) regression effect in a series of magnitude judgments and productions (Cross, 1973).

Local Contraction Bias

In addition to the global effect of the contraction bias in direct magnitude estimation and production that has already been discussed, the contraction bias has a local effect, which is illustrated in Figure 1D. The local contraction bias corresponds to the time–order error that is found in measurements of the differential threshold (Hollingworth, 1910; Woodworth, 1938, pp. 438–448).

When presented in a small range, very high-intensity and very low-intensity stimuli are treated as if they had less extreme values. At the top of Figure 1D for a very intense standard, increasing the subjective intensity by a fixed proportion requires a smaller change in physical intensity than reducing the subjective intensity by the same proportion. At the bottom of the figure for a barely perceptible standard, the tendency is reversed. Increasing the subjective intensity by a fixed proportion requires a larger

change in physical intensity than reducing the subjective intensity by the same proportion.

As yet the local contraction bias appears to have been described only in judgments of loudness in a within-subject experimental design. The bias is illustrated in Figure 8. The figure shows the median number of decibels required to halve and double the loudness of white noise in different 10-dB ranges of sound pressure level, using the method of magnitude production (Poulton & Stevens, 1955). Each of 36 observers sets the intensity of a variable noise to what he judges to be half or double the loudness of a number of standard noises, using a volume control. He pauses for about 1.0 sec between listening to the standard noise and listening to the variable noise. In the range between 90 and 110 dB (SPL), doubling the loudness requires fewer decibels than halving does. While in the range between 30 and 50 dB, doubling requires more decibels than halving does. Over the middle range between 50 and 90 dB there is no obvious asymmetry. This corresponds to the absence of bias in the middle of the range in Figure 1D.

Two obvious explanations of the local contraction bias of Figure 8 can be rejected. Although the contraction occurs at both ends of the range, it is not due to a ceiling or floor effect from the observer approaching the end of his range of responses. At each intensity level the intensity corresponding to the standard is always about one quarter of the way from the lower end of the volume control in doubling and about one quarter of the way from the upper end in halving. Thus, there is always plenty of movement of the volume control available, with the corresponding increase or decrease in intensity.

A second obvious explanation is that the contraction bias with the ranges of intensities centered on 95 and 105 dB could be due to the observer refusing to set the intensity of his volume control high enough. Similarly the contraction bias with the ranges centered on 35 and 45 dB could be due to the observer refusing to set the intensity of his volume control low enough. However, these explanations do not account for the similar biases

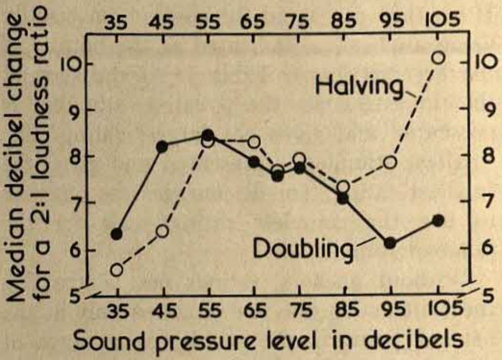

Figure 8. The local contraction bias in magnitude production with small very high-intensity and very low-intensity ranges of stimuli. (Results are from Poulton & Stevens, 1955, Figure 3.)

found with numerical magnitude estimation (S. S. Stevens, 1955b, Tables 1 and 2), because here the experimenter sets the variable sound intensities, and the observer simply judges how much louder or softer they are than the standard.

A local contraction bias similar to the dashed line of Figure 8 for halving loudness is reported by Warren (1970, 1973) for very first judgments of half loudness, for both white noise (Warren, 1973, Figure 2) and a 1-kHz tone (Warren, 1970, Figure 3). In both of Warren's investigations reliably more decibels are required to produce numerical judgments of one half of a sound of 100 dB (SPL) than to produce numerical judgments of one half of a sound of moderate intensity, whereas reliably fewer decibels are required to produce numerical judgments of one half of a sound of 35 dB. For the intermediate ranges between 45 and 90 dB, the number of decibels required for half loudness remains approximately constant.

As in the experiment of Figure 8, in both of Warren's (1970, 1973) experiments the local contraction bias affects the judged subjective intensity of the sound. The bias does not affect the number dimension, whether numbers are used as stimuli or as responses.

Avoiding the Contraction Biases

Investigators using ratings may be concerned only with the relative positions of the sensory magnitudes on their rating scale.

If so, they can avoid the contraction bias by using anchors, as indicated at the bottom of the first section of Table 1. At the start of the investigation the greatest stimulus is presented and given the largest rating. The smallest stimulus is presented and given the smallest rating. This encourages the observer to use the complete rating scale for the range of stimuli.

Without anchors, ratings not affected by the contraction bias are obtained only at the neutral point in the middle of a range of named categories. Here the stimulus is not underestimated because it is not judged to be large. It is not overestimated because it is not judged to be small.

In magnitude estimation, the contraction bias can be avoided in theory by using only the judgments of subjectively moderate differences in intensity. The insuperable difficulty is that the observer is not asked to indicate what represents a subjectively moderate difference in intensity, and there is no other way of specifying it using the experimental data.

S. S. Stevens (1971, p. 444) neatly sidesteps this difficulty by following a suggestion he makes 15 years earlier (S. S. Stevens & Poulton, 1956). The contraction biases that occur in magnitude estimation and in magnitude production can be counterbalanced by averaging the results of the numerical judgments and gain adjustments. Since the contraction biases of the two methods are by definition of equal size, they cancel each other because they are in opposite directions.

In judgments of half magnitude, it is not possible to avoid the contraction bias. Warren (1970, 1973) neatly avoids the bias introduced by the observer's tendency to select a response too close to the middle of the range of responses in very first judgments of half loudness. Warren does this by using only the judgments of the variable that receives an average judgment of exactly half the modulus, or number allocated to the standard. This value is neither overestimated nor underestimated. But unfortunately the method can be used only for judgments of half, which is a relatively small ratio. Small ratios tend to be overestimated. Thus Warren's half-loudness judgments give a steep slope on a loglog plot.

The top of the second section of Table 1 indicates that the local contraction bias of Figure 8 for magnitude judgments with small ranges of very high- or low-intensity stimuli can be avoided only by sticking to stimuli of intermediate intensity, as illustrated in the figure. But this still leaves the main contraction bias, which affects all judgments of small stimulus ratios like half and twice loudness.

Stimulus Equalizing Bias

Figure 1B illustrates a model of the stimulus and the response equalizing biases. For the stimulus equalizing bias the two scales at the sides represent stimuli, whereas the scale in the middle represents responses. The observer uses his full range of responses, whatever the size of the range of stimuli. He simply magnifies his response scale to fit a large stimulus range and shrinks his response scale to fit a small stimulus range (Jones & Woskow, 1962). Like the contraction bias, the stimulus equalizing bias affects the observer's very first judgment, because he always has some idea of the range of stimuli to expect.

Stimulus Equalizing Bias in Rating

Harvey and Campbell (1963) compare two ranges of five weights. Columns 2 and 3 of Table 2 show that Weights A and B of 12.7 and 17.9 ounces (362 g and 508 g, respectively) are numbers 1 and 5 at the ends of the narrow range, but occupy the intermediate serial positions 2 and 4 of the wide range.

The two far-right columns of the table give the mean category judgments of separate groups of 40 undergraduates using a 5-point numerical scale. The bottom row shows that the difference between Weights A and B is 2.5 category steps when they are at the ends of the narrow range but only 1.7 category steps when they occupy the intermediate positions in the wide range. Thus the average category rating received by a weight depends partly on the size of the

Table 2
Stimulus Equalizing Bias in Rating Weights

	Weight in ounces		Judgment in ounces		M category judgment using 5-point scale	
Weight	Wide range	Narrow range	Wide range	Narrow range	Wide range	Narrow range
	10.7					
A	12.7	12.7	10.1	11.4	2.4	2.0
		13.9				
	15.1	15.1	14.7	16.4	3.4	3.3
		16.4				
B	17.9	17.9	21.5	22.8	4.1	4.5
	21.2					
B − A	5.8	5.8	11.4	11.4	1.7[a]	2.5[a]

Note. Each column represents results from a separate group of 40 undergraduates (Harvey & Campbell, 1963).
[a] For wide–narrow range, $p < .01$.

range of weights being judged and only partly on its actual weight.

Parducci and Perrett (1971, p. 436) report a similar effect of the size of the range of stimuli. They obtain judgments of the sizes of squares using nine named categories instead of the five numbered categories of Table 2. The average named category given to a particular square changes by 1.5 category steps when the position of the square is changed from the largest in the range to the third largest.

Judgments in Familiar Physical Units

The two middle columns of Table 2 give results from two more separate groups of 40 undergraduates. They show that the stimulus equalizing bias does not affect the judgments of physical weight. The bottom row shows that the difference of 11.4 ounces (323 g) between the judgments of Weights A and B is the same in both ranges. This is because the response scale of physical weight is closely bound to the stimulus scale of physical weight, which students are taught to use at school. Similarly, in Parducci's (1963) results the stimulus equalizing bias affects the numerical category ratings of the number of dots in a rectangle (Parducci, 1963, Figure 22) but not the judgments of the actual number of dots (Parducci, 1963, Figure 12).

In the display in the section entitled Kinds of Responses at the beginning of the article, the response scales are ordered by the closeness of their links to the stimulus scale. First comes a familiar physical measure of the stimulus such as weight in ounces or the actual number of dots. These measures are not affected by the stimulus equalizing bias because the stimuli and responses are closely linked by well-known rules. The remaining kinds of responses in the display, category ratings and magnitude estimates, can all be affected by the stimulus equalizing bias.

Stimulus Equalizing Bias in Direct Magnitude Estimation

The stimulus equalizing bias can have a marked effect in direct magnitude estimation. The observer starts the investigation with what he believes to be a sensible range of responses. He distributes these responses over the range of stimuli presented to him. Thus the observer who is given the smaller stimulus range produces the steeper slope when the data are presented on a loglog plot (Jones & Woskow, 1962).

This is illustrated by the filled points in Figure 6 (results from Montgomery, 1975). The group of 10 psychology undergraduates represented by the filled triangles judge the loudness of the white noise covering a 25-dB range, from a sound pressure level of 70 dB to 95 dB. The separate group represented by the filled circles has a range of 70 dB, almost

three times as large, from a sound pressure level of 30 dB to 100 dB. The ratio of the slopes of the two fitted straight lines is 2:1.

The observer's judgments are not determined entirely by the size of the range of sound levels heard. If they were, the steeper function in Figure 6 would have almost three times the slope of the less steep function, not twice the slope. But the observer's judgments are determined more by the size of the stimulus range than by the differences in loudness.

There are a number of examples of the stimulus equalizing bias in direct magnitude estimation using separate groups of observers for each stimulus range (Poulton, 1968). The bias has a marked effect on dimensions like loudness and brightness, which students are not taught to handle at school. Using a 1-kHz tone, Engen and Levy (1958, Experiment 2) compare a stimulus range extending from 50 dB to 75 dB above threshold, with a range extending from 25 dB to 75 dB. For separate groups of 10 observers, doubling the stimulus range reduces the average slope for loudness on a loglog plot like that of Figure 6 by a ratio of 1.6:1.

Using white noise, Frederiksen (1975, Experiment 2) compares a stimulus range extending from a sound pressure level of 42 dB to 61 dB with a range extending from 42 dB to 80 dB. For separate groups of 10 undergraduates, doubling the stimulus range reduces the slope by a ratio of 1.6:1. In the corresponding Experiment 1 for light intensity, doubling the physical range reduces the slope by a ratio of 1.7:1.

The stimulus equalizing bias has a smaller effect on dimensions such as length and distance, which students are taught to handle at school. In discussing the stimulus equalizing bias within a sensory dimension, Teghtsoonian (1973) gives only three sets of data, all his own. Two are for apparent distance and apparent length, where the stimulus equalizing bias is small as expected.

The third set of data is for the loudness of a 3-kHz tone. Teghtsoonian presents ranges of sound pressure level centered on 84 dB to separate groups of 16 observers. Here doubling the stimulus range from 20 dB to 40 dB has no effect on the slope of the loudness function. It is not clear why this occurs. The instructions state that "the ratios of successive loudnesses might sometimes be very large or very small" (Teghtsoonian & Teghtsoonian, 1978, p. 307). Perhaps this encourages the observers with the large 40-dB range not to underestimate the ratios and so not to be influenced by the stimulus equalizing bias. The effect of transfer from the instructions is discussed in a later section of this review. Whatever the reason, Teghtsoonian (1973, Figure 4) gives the impression that the stimulus equalizing bias has little effect within a sensory dimension. Yet the results just discussed indicate that this is not usually so for loudness or brightness.

Size of the Stimulus Range in Cross-modal Comparisons

In cross-modal comparisons, the inverse relationship between the exponent or slope on a loglog plot and the size of the stimulus range is first pointed out by Jones and Woskow (1966). They use the data from S. S. Stevens (1960, Figure 11 and Table 3), which show the exponents obtained for a number of sensory dimensions, using force of handgrip for the responses. When the exponents are plotted against the reciprocals of the log geometric stimulus ranges, the straight line fitted to the points accounts for 96% of the variance of the points (Teghtsoonian, 1973).

Figure 9 shows the corresponding relationship for 21 of S. S. Stevens's (Note 3, Table 1) investigations using direct numerical magnitude estimation. S. S. Stevens's data are taken from Poulton's summary (1967, Table 1), but with the exponents calculated from the actual subjective and physical ranges used instead of taking S. S. Stevens's best fitting exponents derived from a number of investigations. Also, the physical ranges of sound and vibration are transformed into units of amplitude instead of power by taking the square root, following S. S. Stevens (1966) and Teghtsoonian (1971). S. S. Stevens uses amplitude instead of power because amplitude is more comparable with the measures of length that he uses for other

sensory dimensions. Following S. S. Stevens (1966), in Figure 9 the square root is also taken of the physical range for light, for the dimensions both of brightness and of the lightness of grays. This is because light and sound behave so similarly (J. C. Stevens & Marks, 1965; S. S. Stevens, 1955a). But the fit is slightly better if the range for light is left uncorrected, as Teghtsoonian (1971) does. The figure follows the method of analysis described by Teghtsoonian. The fitted straight line has a slope of 1.43 or log 27. It accounts for 83% of the variance of the points.

Instead of using the exponents of Figure 9, in Figure 10 the average geometric subjective ranges are plotted against the geometric stimulus ranges on a loglog plot. The figure shows that there is little relationship between the sizes of the subjective ranges and the sizes of the stimulus ranges. The tau coefficient of rank correlation is only .28. For 21 points this is not reliable ($.1 > p > .05$).

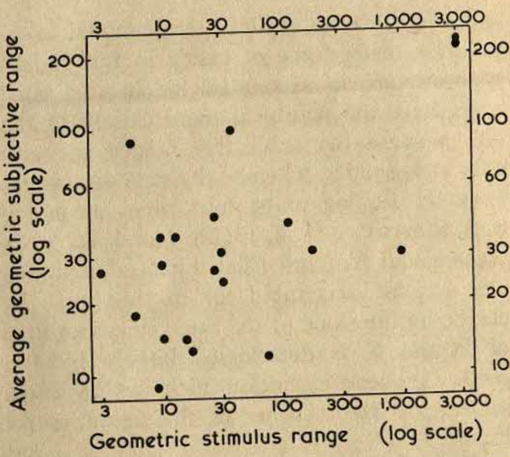

Figure 10. The virtual absence of the stimulus equalizing bias in S. S. Stevens's (Note 3) comparisons between sensory dimensions when exponents are not used.

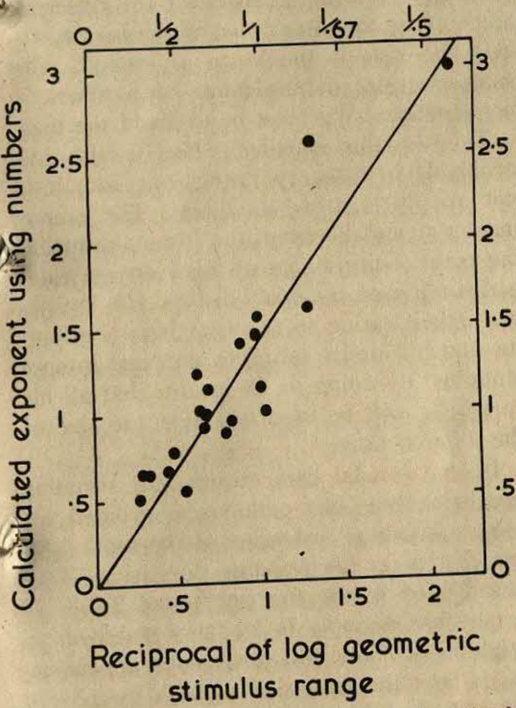

Figure 9. The effect of the size of the stimulus range on the exponents of S. S. Stevens's (Note 3) investigations for different sensory dimensions, using numbers to indicate subjective magnitudes. (The exponents are calculated from the stimulus and response ranges used; they are not precisely the exponents quoted by Stevens. Adapted from Poulton, 1977, Figure 9.)

The overwhelming relationship illustrated in Figure 9 is produced by transforming the subjective ranges into exponents. Figure 10 shows that the subjective ranges run from 9 to 225, a ratio of 25 or 1.4 log units, whereas the stimulus ranges run from 3 to 3,000, a ratio of 1,000 or 3 log units, which is over twice as large in log units. The exponents are the ratios of the log subjective and log stimulus ranges (Teghtsoonian, 1971):

$$\exp = (\log \psi \text{ range})/(\log \phi \text{ range}).$$

Because the subjective ranges vary so much less than the stimulus ranges, they can be approximated by a constant. Using for the constant the slope of the line in Figure 9,

$$\exp = (\log 27)/(\log \phi \text{ range}).$$

The value 27 is the average subjective range of the point for electric shock on the extreme left of Figure 10, which has the smallest stimulus range. In Figure 9 this point stands by itself at the top on the right. Owing to its commanding position it exerts the greatest influence on the slope of the line passing through the origin and fitted to all the points. It insures that the slope will be about log 27.

S. S. Stevens's (1960) cross-modal comparisons using force of handgrip for the responses can be accounted for in the same way. Here the points fit more closely to the line because the subjective ranges vary by only .4 log units, whereas the stimulus ranges vary by 1.6 log units, four times as much. S. S. Stevens's (1966, Figure 2 and Table 1) cross-modal matches using binaural loudness can also be accounted for in this way. In each case the slope of the fitted line, like that of Figure 9, is determined largely by the point for electric shock, which lies by itself in the top right corner of the figure, as in Figure 9.

Clearly Figure 10 is a more appropriate plot for comparing the data from different sensory modalities than a plot like Figure 9. As S. S. Stevens (1971) points out, his stimulus ranges usually extend from not far above threshold to near the maximum that the investigator can obtain or the observer will tolerate. Thus, on both axes of Figure 10 the points are reasonably representative. They show how most of the observer's subjective range on each dimension is related to most of his stimulus range.

Perhaps in the future those interested in comparing sensory magnitudes across modalities will describe the relationships as they are presented in Figure 10. Thus both loudness and brightness have an average subjective range of about 220:1 for a physical amplitude (power$^{\frac{1}{2}}$) range of about 3,000:1. Electric shock has an average subjective range of about 27:1 for a physical-range of about 3:1, whereas cold on the front of the forearm has both an average subjective range and a physical range of about 9:1.

However, Teghtsoonian (1971, 1973) uses the fit of the line like that in Figure 9 and the fits of the corresponding lines with force of handgrip and binaural loudness as the dependent variables, as evidence that S. S. Stevens's subjective ranges are "nearly constant." This conclusion is not compatible with S. S. Stevens's (Note 3) data in Figure 10, nor is it compatible with the comparable data using force of handgrip (S. S. Stevens, 1960) and binaural loudness (S. S. Stevens, 1966). The excellent fits of the lines indicate only that the stimulus ranges are very much larger than the subjective ranges.

Teghtsoonian (1973) also states that variations in the stimulus range have a marked effect in comparisons across modalities but little effect within modalities. This is the wrong way around. The stimulus range has no reliable effect across modalities (Figure 10), unless comparisons are made between exponents. But the stimulus range has a marked effect within modalities that students are not used to handling, such as loudness and brightness.

Avoiding the Stimulus Equalizing Bias

The stimulus equalizing bias can be avoided only by using a range of stimuli of the same subjective size as the range of responses. The filled points in Figure 6 show that the bias cannot be avoided in direct magnitude estimation, because the observer tends to use the same range of numbers whatever the size of the range of stimuli. The subjective size of the range of numbers is not constant, as it must be to avoid the bias.

The stimulus equalizing bias is also unavoidable in category rating, but here it is not usually regarded as a bias. The experimenter may make use of the stimulus equalizing bias to insure that all his response categories are used reasonably often. He anchors his highest rating to the most intense stimulus and his lowest rating to the least intense stimulus. By doing so, he insures that all his categories will be used whatever the size of the stimulus range.

In cross-modal comparisons, the stimulus equalizing bias can perhaps be avoided by using a stimulus dimension of the same subjective size as the response dimension. This is indicated in the first section of Table 1. A possible example is loudness matched to brightness. These two dimensions appear to have about the same sizes of both the physical and the subjective ranges (J. C. Stevens & Marks, 1965; S. S. Stevens, 1955a).

Response Equalizing Bias

Figure 1B illustrates the response equalizing bias as well as the stimulus equalizing

bias. For the response equalizing bias the two scales at the sides represent responses, while the scale in the middle represents stimuli. Whatever range of responses the observer is given, he distributes the responses over the range of stimuli. He uses a larger range of responses when a larger range is available.

Response Equalizing Bias in Rating

The response equalizing bias is not usually regarded as a bias when rating with a small number of categories. In rating sensory magnitudes, the investigator often encourages the response equalizing bias by his instruction. He tells his observer to use the full range of categories that he is given. The instruction insures that the proportional relationships between a set of variables will hardly be affected by the exact number of categories used. This applies to small sets of both named and numbered categories. Thus on a linear plot there is a nearly perfect linear relationship between the average named ratings on a 6-point scale and the average named ratings from separate groups of observers using a 9-point scale (Parducci & Perrett, 1971, Figure 4).

The linear relationship does not hold between a small set of categories and 50 or 100 categories, because the use of both single-digit and two-digit categories introduces the logarithmic bias of Figure 1F.

Response Equalizing Bias in Very First Judgments of Reflectance

In direct magnitude estimation, the response equalizing bias can be demonstrated by giving a standard stimulus a modulus of 10. Separate groups of observers judge the same variable against the same standard, using numbers ranging either between 10 and 0 or between 10 and infinity. Poulton and Simmonds (1963) compare the two response ranges in judgments of the reflectance of gray papers. They use 10 pairs of groups, each group comprising 50 students making their very first magnitude judgments. Of the 10 groups that use numbers ranging between 10 and infinity, 9 have median judgments that differ from the standard by at least 10, whereas none of the 10 groups that judge the same stimuli using numbers ranging between 10 and 0 have median judgments that differ from the standard by as much as 10. To do so, the median would have to be 0. Thus, a larger range of numbers is used when a larger range is available.

Avoiding the Response Equalizing Bias

The response equalizing bias can be avoided in direct magnitude estimation by leaving the observer free to choose his own range of responses. The bias is unavoidable in category rating. As already indicated, in category rating the investigator may encourage the response equalizing bias. It makes his proportional relationships almost independent of the exact number of categories that he uses, provided he sticks to nine categories or less.

In cross-modal matching, the investigator can avoid introducing the response equalizing bias by using a response dimension of the same subjective size as the stimulus dimension. This is indicated in the first section of Table 1.

Bias From Transfer

The biases that result from transfer can be classified by what is transferred:

1. *Transfer from previous stimuli:* Here the present stimulus is judged in the context of previous stimuli. Examples are the centering bias and the stimulus spacing bias of Figures 1A and 1E. These two transfer biases occur gradually as the observer learns the set of stimuli used in the investigation. Another example is the sequential contraction bias, which depends on the immediately preceding stimulus (Jesteadt et al., 1977). These transfer biases have already been discussed. In this section we consider only the remaining transfer biases.

2. *Transfer from previous judgments:* Here it is not just the previous stimuli but also the observer's responses to them that bias his judgments. The present response is made to be consistent with the responses made previously to previous stimuli. Examples of the

influence of one judgment on the next are the transfer between ranges of stimuli, the transfer between ranges of responses, and the transfer from interval to ratio judgments.

3. *Transfer from previous responses:* Here the observer tends to use the same range of responses as he uses in a previous condition with different stimuli.

Transfer Between Ranges of Stimuli

In investigations on transfer between ranges of stimuli, the stimuli to be judged are increased in magnitude or reduced in magnitude. Compared with the judgments of a control group that does not experience the change, after a range of large stimuli, smaller stimuli are judged too small. After a range of small stimuli, larger stimuli are judged too large.

The transfer can be explained as follows: The centering bias of Figure 1A affects the observer's responses to the first range of stimuli. It leaves only the more extreme responses available for the second range of stimuli. The stimuli are therefore judged more extreme than they should be.

Transfer occurs even when the stimuli are judged in familiar physical units such as length in inches, provided the stimulus lines are flashed on a screen for only .2 sec and are therefore not easy to judge accurately (Krantz & Campbell, 1961). Transfer occurs when the stimuli are judged in named categories (DiLollo, 1964; Parducci, 1954, 1956; Ross & DiLollo, 1968a; Tresselt, 1947). Transfer also occurs when a small set of numbered categories is used (Campbell, Lewis, & Hunt, 1958; Melamed & Thurlow, 1971; Parducci, 1963; Pollack, 1964b, Experiment 1), when a potentially large set of numbers is used centered on 100 (Krantz & Campbell, 1961) or extending up to 100 (DiLollo & Kirkham, 1969), and when an infinite set of numbers is used, as in direct numerical magnitude estimation (Melamed & Thurlow, 1971; Ross & DiLollo, 1968b).

Transfer Between Ranges of Responses

The response equalizing bias in very first judgments of reflectance has just been described (Poulton & Simmonds, 1963). Here there is marked transfer from the very first judgments using finite (10 to 0) or infinite (10 to infinity) ranges of numbers to second judgments of the other kind using an infinite range after a finite range or a finite range after an infinite range.

For the very first judgments, 8 out of the 10 comparisons between the finite and the corresponding infinite ranges show that a reliably larger range of numbers is used when the larger range is available. However, on transfer to the second judgment of the opposite kind, 7 of the 8 possible matched comparisons before and after transfer are less marked after transfer ($p < .05$ on a two-tailed Wilcoxon test). Only 3 of the 8 comparisons after transfer still show that a reliably larger range of numbers is used when the larger range is available (Poulton, 1968, Figure 2 and p. 10). Thus after a judgment using the other range of numbers, finite and infinite ranges of numbers give more similar judgments than they do when they are used for very first judgments.

This is presumably because for the very first judgments the students consider only a range of numbers extending from the modulus of 10 either to 0 or to infinity, whereas the second judgments of both kinds are made in the context of a total range of numbers extending from 0 to infinity. The second judgments are therefore more similar than the first judgments.

Transfer From Interval to Ratio Judgments

Fagot, Eskildsen, and Stewart (1966) report reliable transfer from a set of interval judgments made over 3 successive days to a set of ratio judgments made on 3 subsequent successive days (Fagot, Note 4). For the interval or bisection judgments the observer controls the brightness of a circle of light placed between two standard circles of 580 fL and 2 fL, respectively. For the ratio or fractionation judgments, the observer does not see the circle of 2 fL. He adjusts the brightness of the light that he controls to half the brightness of the light of 580 fL.

When the ratio judgments are performed first, the average geometric mean for half

brightness is 52 fL, giving a mean slope on a loglog plot of .28. When the ratio judgments are performed after the interval judgments, half brightness falls reliably to 13 fL, giving a mean slope of only .18. The ratio of the two slopes is 1.6:1. The interval bisection judgments are not appreciably affected by prior ratio judgments. The average geometric mean falls only from 47 to 41 fL. This could be due to initial differences between the two groups of four observers who perform the two conditions in opposite orders.

The reliable transfer can be explained quite simply in terms of the hypothetical zero used for the ratio judgments. Suppose an observer first gives an average bisection judgment of 47 fL between the standards of 580 and 2 fL. When the standard of 2 fL is removed for the ratio judgments, the observer realizes that he has now to judge a larger range of brightnesses, extending from 580 fL to 0 fL. He therefore selects a half brightness less than his previous average of 47 fL.

In contrast, the observer who starts by making the ratio judgments has to get on as best he can without having seen a specified lower bound to the range of brightnesses. When he is subsequently presented with a lower bound for his bisection judgments, he of course uses it.

Transfer From the Range of Numbers Used in the Instructions

In direct magnitude estimation, the numerical examples used by the investigator in explaining the technique can reliably affect the slope of the psychophysical function that he obtains. In a series of investigations, G. H. Robinson (1976) gives alternate observers different numerical examples. The standard stimulus is always called 100. For half the observers the written instructions state that if the variable is one and a half times the standard, it should be called 150. If the variable is half the standard, it should be called 50. For the other half of the observers the instructions state that if the variable is seven and a half times the standard it should be called 750. If the variable is one quarter of the standard, it should be called 25. The range of numbers specified is simply increased from 50–150 to 25–750, an increase in ratio from 1:3 to 1:30.

In G. H. Robinson's (1976) Experiment 1 for magnitude estimations of auditory pulse rate, the increase in the specified range of numbers increases the exponent from .87 to 1.3 and from .89 to 1.3 in a replication, ratios of about 1:1.5. In Experiment 2 for the loudness of a 1-kHz square wave, the exponent increases from .73 to 1.0 and from .77 to .95 in a replication, ratios of about 1:1.3. All four increases are reliable. The plotted functions are concave downward like those at the top of Figure 3B. Each function for the larger range of numbers is a little steeper than the function for the smaller range. For the smaller range of numbers, each function is a good deal more concave downward at the upper end than the functions at the top of Figure 3B, as if the observers run out of high numbers.

Clearly, the method of direct magnitude estimation is extremely sensitive to numerical examples given in the instructions. If an investigator wishes to obtain unbiased data, he should not use numerical examples. Unfortunately S. S. Stevens (1971, p. 428) recommends their use, suggesting numbers 20 times and one fifth of the standard, a ratio of 1:100. Only 4 of Stevens's 21 average geometric subjective ranges, illustrated in Figure 10, are as large or larger than this. So if he does use these instructions, in most of his investigations he is biasing his observers into giving unduly large exponents. This may explain why other investigators often find rather smaller exponents than S. S. Stevens does.

The disturbing feature is that the investigators who now follow Stevens in giving numerical examples may suggest ranges of numbers to their observers that will produce the sizes of exponents they expect to find. Investigators would be foolish to suggest ranges of numbers that produce unexpected sizes of exponents. Yet suggesting appropriate ranges of numbers will produce a spurious consistency between the exponents and the theory. Unfortunately investigators do not often report the exact instructions that

they give. In these cases the reader has no idea whether the investigator finds the size of exponent that he does find simply because he suggests the particular range of numbers to his observers.

Transfer From the Range of Numbers Used in a Previous Investigation

S. S. Stevens and his colleagues use the same members of the departmental staff and graduate students for more than one investigation (S. S. Stevens, 1959). From the evidence just reviewed on the transfer of ranges of numbers, it is almost certain that there is transfer of the range of numbers from one investigation to the next. This is a bias that S. S. Stevens and his students ignore.

S. S. Stevens (1971, p. 428) even recommends "under some circumstances," which he does not specify, clarifying the nature of the task before starting an investigation. This is done by getting the observer to match numbers "to an easier continuum, such as apparent length of lines, or apparent size of circles" (Stevens, 1971, p. 428). Yet as is pointed out in an earlier review (Poulton, 1968),

> It would be too easy for the experimenter to select a range of lengths of lines calculated to elicit the set of numbers which he believed to be appropriate to the range of stimuli which he was subsequently going to present to the observer. Clearly, the experimenter would be foolish to present lines which elicit a quite inappropriate set of numbers, in view of the expected transfer effects. The experimenter is thus in a predicament which is better avoided. (p. 4)

Avoiding Bias From Transfer

As the third section of Table 1 indicates, the only way of avoiding bias from prior conditions is to use separate groups of observers for each condition. Bias from instructions and demonstrations can be avoided by using unbiased instructions and no demonstrations. Transfer from previous stimuli, judgments, and responses can be avoided only by restricting each group of observers to a single judgment.

The difficulty with using separate groups is the large individual differences in judging sensory magnitudes. S. S. Stevens (1971, Figure 3) reports that his individual exponents for loudness range from .4 to 1.1, a ratio of 1:2.7. To obtain a reasonably precise average numerical value, it is therefore necessary to use large numbers of observers. Yet S. S. Stevens and his colleagues typically use only 10 or 12 observers in each investigation. To obtain data that are comparable across investigations, they need to stick to the same observers or to use groups of observers who are known to give comparable magnitude judgments because they have served in previous investigations. Yet the use of selected trained observers almost certainly introduces bias from transfer.

Avoiding all Biases

The fourth section of Table 1 indicates how to avoid all the biases that can be avoided. As already pointed out, in category rating it is not possible to avoid the stimulus and response equalizing biases. An investigator who wishes to avoid these two biases does not use ratings.

In magnitude estimation the response equalizing bias can be avoided by leaving the observer free to choose his own range of responses. However, the only way of avoiding the stimulus equalizing bias is by the cross-modal matching of sensory dimensions with subjectively equal-sized ranges, as indicated at the bottom of Table 1 on the right. This matches only subjective magnitudes in one sensory dimension to subjective magnitudes in another sensory dimension with a subjectively equal-sized range. There is no sure unbiased method of matching subjective magnitudes to numbers, because the observer tends to use the same range of numbers whatever the size of the range of stimuli. Thus the subjective size of the range of numbers varies with the size of the range of stimuli.

Clearly, avoiding all possible biases provides so little useful data for so much effort that not many investigators are likely to contemplate it. The most practical alternative is to collect data with the minimum of known biases by balancing the biases against each other, as suggested by S. S. Stevens (1971, pp. 444–446). Then make checks using

separate groups of untrained observers to estimate the sizes of the residual biases. Finally, the data should be corrected for the residual biases.

A difficulty with this alternative is that the effects of the residual biases may be considerably larger than the effects that the investigator wishes to study. Also, the biases may interact with the effects being studied. Unfortunately, at the present time most investigators simply collect biased data without attempting to correct for the biases or even to measure and report them.

Reference Notes

1. Lauber, A. *Schallmessungen an Motorfahrzeugen mit subjektivem Hörvergleich* (Report No. 22-637). Bern, Switzerland: Generaldirektion PTT, Forschungs und Versuchsanstalt, October 1959.
2. Joynson, R. B. Personal communication, January 26, 1978.
3. Stevens, S. S. *In pursuit of the sensory law* (2nd Klopsteg lecture). Evanston, Ill.: Northwestern University, Technological Institute, November 1962.
4. Fagot, R. F. Personal communication, October 26, 1977.

References

Anderson, N. H. Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press, 1974.

Andrews, B., & Finch, D. M. Truck-noise measurement. *Proceedings of the Highway Research Board*, 1951, *31*, 456–465.

Banks, W. P., & Hill, D. K. The apparent magnitude of number scaled by random production. *Journal of Experimental Psychology*, 1974, *102*, 353–376.

Bowsher, J. M., Johnson, D. R., & Robinson, D. W. A further experiment on judging the noisiness of aircraft in flight. *Acustica*, 1966, *17*, 245–267.

Campbell, D. T., Lewis, N. A., & Hunt, W. A. Context effects with judgmental language that is absolute, extensive, and extra-experimentally anchored. *Journal of Experimental Psychology*, 1958, *55*, 220–228.

Cross, D. V. Sequential dependencies and regression in psychophysical judgments. *Perception & Psychophysics*, 1973, *14*, 547–552.

DiLollo, V. Contrast effects in the judgment of lifted weights. *Journal of Experimental Psychology*, 1964, *68*, 383–387.

DiLollo, V., & Kirkham, R. Judgmental contrast effects in relation to range of stimulus values. *Journal of Experimental Psychology*, 1969, *81*, 421–427.

Eisler, H. Empirical test of a model relating magnitude and category scales. *Scandinavian Journal of Psychology*, 1962, *3*, 88–96.

Engen, T., & Levy, N. The influence of context on constant-sum loudness-judgments. *American Journal of Psychology*, 1958, *71*, 731–736.

Erlebacher, A. Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. *Psychological Bulletin*, 1977, *84*, 212–219.

Fagot, R. F., Eskildsen, P. R., & Stewart, M. R. Effect of rate of change in physical intensity on bisection and fractionation judgments of brightness. *Journal of Experimental Psychology*, 1966, *72*, 880–886.

Frederiksen, J. R. Two models for psychological judgment: Scale invariance with changes in stimulus range. *Perception & Psychophysics*, 1975, *17*, 147–157.

Gibson, R. H., & Tomko, D. L. The relation between category and magnitude estimates of tactile intensity. *Perception & Psychophysics*, 1972, *12*, 135–138.

Greenwald, A. G. Within-subjects designs: To use or not to use? *Psychological Bulletin*, 1976, *83*, 314–320.

Harvey, O. J., & Campbell, D. T. Judgments of weight as affected by adaptation range, adaptation duration, magnitude of unlabeled anchor, and judgmental language. *Journal of Experimental Psychology*, 1963, *65*, 12–21.

Helson, H. *Adaptation-level theory*. New York: Harper & Row, 1964.

Holland, M. K., & Lockhead, G. R. Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, 1968, *3*, 409–414.

Hollingworth, H. L. The central tendency of judgment. *Journal of Philosophy, Psychology, and Scientific Methods*, 1910, *7*, 461–469.

Jesteadt, W., Luce, D. R., & Green, D. M. Sequential effects in judgments of loudness. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 92–104.

Jones, F. N., & Woskow, M. J. On the relationship between estimates of magnitude of loudness and pitch. *American Journal of Psychology*, 1962, *75*, 669–671.

Jones, F. N., & Woskow, M. J. Some effects of context on the slope in magnitude estimation. *Journal of Experimental Psychology*, 1966, 71, 170–176.

Joynson, R. B., Newson, L. J., & May, D. S. The limits of over-constancy. *Quarterly Journal of Experimental Psychology*, 1965, *17*, 209–216.

Krantz, D. L., & Campbell, D. T. Separating perceptual and linguistic effects of context shifts upon absolute judgments. *Journal of Experimental Psychology*, 1961, *62*, 35–42.

Melamed, L. E., & Thurlow, W. R. Analysis of contrast effects in loudness judgments. *Journal of Experimental Psychology*, 1971, *90*, 268–274.

Montgomery, H. Direct estimation: Effect of methodological factors on scale type. *Scandinavian Journal of Psychology*, 1975, *16*, 19–29.

Parducci, A. Learning variables in the judgment of single stimuli. *Journal of Experimental Psychology*, 1954, *48*, 24–30.

Parducci, A. Direction in shift in the judgment of single stimuli. *Journal of Experimental Psychology*, 1956, *51*, 169–178.

Parducci, A. Range-frequency compromise in judgment. *Psychological Monographs*, 1963, *77*(2, Whole No. 565).

Parducci, A., & Perrett, L. F. Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology*, 1971, *89*, 427–452. (Monograph)

Pollack, I. Neutralization of stimulus bias in auditory rating scales. *Journal of the Acoustical Society of America*, 1964, *36*, 1272–1276. (a)

Pollack, I. Order effects in category scaling of grays. *Psychonomic Science*, 1964, *1*, 69–70. (b)

Pollack, I. Iterative techniques for unbiased rating scales. *Quarterly Journal of Experimental Psychology*, 1965, *17*, 139–148. (a)

Pollack, I. Neutralization of stimulus bias in the rating of grays. *Journal of Experimental Psychology*, 1965, *69*, 564–578. (b)

Poulton, E. C. Population norms of top sensory magnitudes and S. S. Stevens' exponents. *Perception & Psychophysics*, 1967, *2*, 312–316.

Poulton, E. C. The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, 1968, *69*, 1–19.

Poulton, E. C. Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin*, 1973, *80*, 113–121.

Poulton, E. C. *Tracking skill and manual control*. New York: Academic Press, 1974.

Poulton, E. C. Range effects in experiments with people. *American Journal of Psychology*, 1975, *88*, 3–32.

Poulton, E. C. Quantitative subjective assessments are almost always biased, sometimes completely misleading. *British Journal of Psychology*, 1977, *68*, 409–425.

Poulton, E. C., & Freeman, P. R. Unwanted asymmetrical transfer effects with balanced experimental designs. *Psychological Bulletin*, 1966, *66*, 1–8.

Poulton, E. C., & Simmonds, D. C. V. Value of standard and very first variable in judgments of reflectance of grays with various ranges of available numbers. *Journal of Experimental Psychology*, 1963, *65*, 297–304.

Poulton, E. C., & Stevens, S. S. On the halving and doubling of the loudness of white noise. *Journal of the Acoustical Society of America*, 1955, *27*, 329–331.

Robinson, D. W., Copeland, W. C., & Rennie, A. J. Motor vehicle noise measurement. *Engineering*, 1961, *211*, 493–497.

Robinson, G. H. Biasing power law exponents by magnitude estimation instructions. *Perception & Psychophysics*, 1976, *19*, 80–84.

Ross, J., & DiLollo, V. Category scales and contrast effects with lifted weights. *Journal of Experimental Psychology*, 1968, *78*, 547–550. (a)

Ross, J., & DiLollo, V. A vector model for psychophysical judgment. *Journal of Experimental Psychology Monograph*, 1968, *77*(3, Pt. 2). (b)

Stevens, J. C. Stimulus spacing and the judgment of loudness. *Journal of Experimental Psychology*, 1958, *56*, 246–250.

Stevens, J. C., & Marks, L. E. Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences*, 1965, *54*, 407–411.

Stevens, J. C., & Tulving, E. Estimations of loudness by a group of untrained observers. *American Journal of Psychology*, 1957, *70*, 600–605.

Stevens, S. S. Decibels of light and sound. *Physics Today*, 1955, *8*, 12–17. (a)

Stevens, S. S. The measurement of loudness. *Journal of the Acoustical Society of America*, 1955, *27*, 815–829. (b)

Stevens, S. S. Cross-modality validation of subjective scales for loudness, vibration, and electric shock. *Journal of Experimental Psychology*, 1959, *57*, 201–209.

Stevens, S. S. The psychophysics of sensory function. *American Scientist*, 1960, *48*, 226–253.

Stevens, S. S. Matching functions between loudness and ten other continua. *Perception & Psychophysics*, 1966, *1*, 5–8.

Stevens, S. S. Issues in psychophysical measurement. *Psychological Bulletin*, 1971, *78*, 426–450.

Stevens, S. S. *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley, 1975.

Stevens, S. S., & Galanter, E. H. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 1957, *54*, 377–411.

Stevens, S. S., & Greenbaum, H. B. Regression effect in psychophysical judgment. *Perception & Psychophysics*, 1966, *1*, 439–446.

Stevens, S. S., & Poulton, E. C. The estimation of loudness by unpracticed observers. *Journal of Experimental Psychology*, 1956, *51*, 71–78.

Teghtsoonian, R. On the exponents in Stevens' law and the constant in Ekman's law. *Psychological Review*, 1971, *78*, 71–80.

Teghtsoonian, R. Range effects in psychophysical scaling and a revision of Stevens' law. *American Journal of Psychology*, 1973, *86*, 3–27.

Teghtsoonian, R., & Teghtsoonian, M. Range and regression effects in magnitude scaling. *Perception & Psychophysics*, 1978, *24*, 305–314.

Torgerson, W. S. Quantitative judgment scales. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications*. New York: Wiley, 1960.

Tresselt, M. E. The influence of amount of practice upon the formation of a scale of judgment.

Journal of Experimental Psychology, 1947, *37*, 251–260.

Von Wright, J. M., & Kekkinen, R. Stimulus range and the estimated ratio between two stimuli. *Perceptual and Motor Skills*, 1970, *31*, 294.

Ward, L. M. Category judgments of loudness in the absence of an experimenter-induced identification function: Sequential effects and power-function fit. *Journal of Experimental Psychology*, 1972, *94*, 179–184.

Ward, L. M. Repeated magnitude estimations with a variable standard: Sequential effects and other properties. *Perception & Psychophysics*, 1973, *13*, 193–200.

Ward, L. M. Sequential dependencies and response range in cross-modality matches of duration to loudness. *Perception & Psychophysics*, 1975, *18*, 217–223.

Ward, L. M., & Lockhead, G. R. Response system processes in absolute judgment. *Perception & Psychophysics*, 1971, *9*, 73–78.

Warren, R. M. Elimination of biases in loudness judgments for tones. *Journal of the Acoustical Society of America*, 1970, *48*, 1397–1403.

Warren, R. M. Quantification of loudness. *American Journal of Psychology*, 1973, *86*, 807–825.

Woodworth, R. S. *Experimental psychology*. New York: Holt, 1938.

Received March 21, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 804–830

Behavioral Treatment of Children's Fears: A Review

Anthony M. Graziano, Ina Sue DeGiovanni, and Kathleen A. Garcia
State University of New York at Buffalo

Behavioral literature on childhood fears, including conceptual models, normative research, and fear-reduction studies is reviewed. The main conclusions are as follows: (a) The information value of nearly 60 years of normative studies is meager, and their continuation is of doubtful value; (b) most research has been limited to laboratory studies of mildly to moderately fearful children, and few data exist on severe fears studied in the child's natural environment or on the clinical prevalence of fear; (c) cognitive and developmental factors have been largely ignored; (d) modeling is the most frequently used and reliably effective fear-reduction strategy; (e) a cognitive, verbal–mediation approach is promising, but is not yet sufficiently researched; (f) there is little evidence that systematic desensitization or contingency management strategies are effective. Implications for large-scale fear reduction and prevention are discussed. The need for research that recognizes the complex paradigms of children's fears is suggested.

This article is a selective review of behavioral treatment of children's fears, an area relatively neglected by behavior therapists and researchers despite their considerable attention to adult fear reduction (Graziano, 1975). Adults seem to minimize the importance of children's fears, viewing them as common and transitory and thus not a particularly serious part of normal development. But children's fears may not always be transient, and some, such as specific animal phobias (Jersild, 1968; Jersild & Holmes, 1935a; Marks & Gelder, 1966) and fear of physical injury or psychic stress (L. C. Miller, Barrett, Hampe, & Noble, 1972b) may persist as adult problems. Children do experience fears, often intense and disturbing, and the psychological suffering of a fearful child, even if it remits in a few years, is at least as worthy of professional concern as is the suffering of adults. There seems good reason for urging more study of fear reduction in children.

Because of this review's behavioral focus, the large psychoanalytic literature is not included. Likewise, school phobia articles, which outnumber those of any other fears by about 25:1 (L. C. Miller, Barrett, & Hampe, 1974), have been well reviewed elsewhere (Gelfand, 1978; Hersen, 1971; Kelly, 1973) and are only briefly discussed.

The modern psychological study of children's fears is nearly 60 years old. The literature consists of psychoanalytic and behavioral case studies, normative fear surveys, controlled fear-reduction experiments that are mainly clinical analogs, and a few theoretical articles and reviews. Much of the literature is psychoanalytic, but interestingly, the earlier (e.g., Jones, 1924a, 1924b; Watson & Rayner, 1920) and the most recent articles have been behavioral.

Berecz (1968) concluded that the research has been "disappointing," giving "only hints" of the nature and incidence of childhood fears. A decade later, the situation regarding normative research appears much the same. A recent increase in behavioral studies of child fear reduction includes some well-designed research (e.g., Bandura & Menlove, 1968; Kanfer, Karoly, & Newman, 1975; Kornhaber & Schroeder, 1975; Melamed & Siegel, 1975; Murphy & Bootzin, 1973). Overall, however, the bulk of research still

Requests for reprints should be sent to Anthony M. Graziano, Department of Psychology, State University of New York, 4230 Ridge Lea Road, Buffalo, New York 14226.

leaves us with hard-to-interpret data, results that are more suggestive than definitive, and an important area that is very much in need of more empirical research.

One reason for the persistence of uncertain results may be a generally held assumption that knowledge about adult fears is sufficient for understanding fears of children, an assumption that may have encouraged inappropriate concepts and approaches. Piaget's work (e.g., 1929, 1951) on cognitive development should alert one to the dangers of considering children to be miniature adults. Berecz (1968) suggested other reasons, namely, a too inclusive use of the term *phobia* to describe "such a variety of conditions" as to make the term meaningless and definitions of phobias that are heavily biased by theoretical conceptions.

Ethical restraints, too, make this a difficult area in which to conduct research. For obvious ethical reasons children cannot be subjected to controlled presentations of fear stimuli in order to measure their reactions. Less intrusive measures, such as self-reports, questionnaires and observations taken in clinical treatment and research, must be used. To these constraints add the problems of obtaining reliable and valid personal information from children, and some of the difficulties become apparent. Accordingly, researchers have tended to tread lightly, using nonintrusive methods that rely heavily on retrospective, subjective reports based on unsystematic self-observations. Vitually all of the normative data were drawn from one form or another of retrospective, subjective report. Frequently the data were from second-person (parents and teachers) verbal reports or were limited to observations made in clinical case studies. These limitations may largely account for the lack of hard data that still characterizes the field despite more than a half century of psychological study.

Definitions: Fears, Phobias, and Clinical Problems

Despite these vagaries, there is fairly good general agreement on definitions of *fear* and *phobia*. Fear is commonly thought of as a normal reaction to genuine threat that involves at least three response systems: (a) overt behavioral expressions, (b) covert, subjective feelings and thoughts, and (c) physiological activity. The term *phobia* is commonly used to denote a specific fear in which at least one of the three elements is excessive, persistent, and unadaptive (Marks, 1969; L. C. Miller et al., 1974). Thus it is generally agreed that fear is a normal response to threatening stimuli, whereas phobia is an unreasonable response, often to usually benign or ill-defined stimuli.

Severity ranges from the mildly fearful responses of many children toward darkness, dogs, and so on to disabling, highly intense levels of fear. It is reasonable to assume that fears presented at clinics for professional intervention are the most severe fears. Two questions are raised immediately when one considers *clinical* fears. First, how are clinical fears to be defined other than by virtue of the fact that clients show up in clinics for treatment? Hampe, Noble, Miller, and Barrett (1973) found that most children, even those with intense fears, overcome them with or without treatment within 2 years. But some do not, and they might appropriately be classed as having fears of *clinical* duration. Intensity and duration might be important defining characteristics, and we suggest that clinical fears be defined as those with a duration of over 2 years or an intensity that is debilitating to the client's routine life-style. The second question is, how prevalent are severe fears and should current research be focused only on severe fears or on both mild and clinical levels of the problem? As is shown in the review of controlled studies, most researchers treat all levels of severity the same, working primarily with mildly fearful children and generalizing to all phobic children. A recent review of behavioral treatment of clinical phobias by Mathews (1978) emphasizes the inappropriateness of generalizing from what are essentially analog studies to treatment of clinical fears. This limitation seems to have occurred because both mild and severe levels of fear are important research areas, and mildly fearful children are considerably easier to recruit as research subjects than are severely fearful children. However, it seems important to identify clearly

the severity of subjects' fears and the limitations that severity places on generalization of the data, especially since one does not know whether mild and clinical fears operate according to identical principles.

We know surprisingly little about childhood fears of high intensity, long duration, and disturbing content. L. C. Miller et al. (1974) estimated that school phobics make up 1% of the school-age population, whereas other child phobics may "run as high as 20 per cent" (p. 91). In his review, Marks (1969) found little data on the frequency with which phobic children are referred for psychiatric treatment. Graham (cited in Marks, 1969) reported five cases of school refusal out of 162 clinical referrals and only 10 specific phobias in 239 cases referred to a children's psychiatric hospital. Rutter, Tizard, and Whitmore (1970) screened all 10- and 11-year-old children ($N = 2{,}199$) on the Isle of Wight and found only 16 children to have clinically significant and handicapping phobias. Their screening procedure omitted those with monosymptomatic phobias, and thus 16 is a conservative estimate. In the most recent attempt to assess clinical prevalence, Graziano and DeGiovanni (1979) obtained questionnaire data from 19 behavior therapists. They reported a total of 547 children referred within the previous 6 months, of which 37, or 6.8%, were referred for treatment of clinical level fears, as defined above. At present only guesses can be made about the nature and extent of the clinical problem of children's fears.

Fear Paradigms

Although several theoretical models have been advanced, none adequately explains the etiology and maintenance of phobias. The major conceptual models are psychoanalytic and behavioral, with ethological theory and a cognitive–developmental model representing less common but potentially important approaches. The major focus of this article is on the behavioral literature.

Behavioral approaches consist of three paradigms, each of which assumes that fears and phobias are learned: respondent conditioning, operant conditioning, and the two-factor theory of learning. The respondent model postulates that anxiety is the central aspect of phobic behavior and that any neutral stimulus present at the time a fear response occurs may become a conditioned stimulus. In future presentations it will elicit the associated fear response. The conditioning occurs if the original conditioning situation was of high intensity or was repeated a number of times. The conditioned fear is presumably maintained in the absence of reinforcement because the avoidance behavior precludes the requirements for extinction, that is, the repeated confrontation of the actually benign conditioned stimulus without its pairing with the unconditioned stimulus. A decrease in avoidance behavior can be achieved by replacing the conditioned anxiety response with a response antagonistic to anxiety (e.g., relaxation) that inhibits or weakens the anxiety. Wolpe's (1974) systematic desensitization is best known of the therapy procedures based on this *reciprocal inhibition* paradigm. The major problems with the respondent model are that it does not explain why some neutral stimuli seem more likely than others to become conditioned fear stimuli, and, although it attempts to, it does not adequately explain the maintenance of phobic behavior in the absence of reinforcement.

Operant models hold that reinforcement, rather than anxiety, primarily social reinforcement such as parental attention, is the central aspect of phobic behavior. Children are presumably taught to be afraid by parents and other significant persons who selectively, albeit unintentionally, attend to and reward fearful behavior. Therapy procedures based on operant models attempt to reduce children's phobic behavior by changing the reward structure in their immediate environments, that is, by teaching parents to ignore phobic behavior and to differentially reward alternative nonfearful approach behaviors. The major problem with the operant paradigm is its failure to adequately account for and therapeutically treat the subjective feelings of intense fear or anxiety and their accompanying thoughts that are often experienced by phobic children.

Modeling, one of the most frequent and promising therapeutic techniques, is based on social learning theory, which is an extension of operant conditioning. Modeling assumes

that behavior can be changed through the vicarious experience of rewards (either tangible or cognitive) for new behavior or of the informational value of the modeled behavior. For example, a child's phobic avoidance of dogs may be decreased by his or her vicarious experience of the rewards given to another child (the model) who approaches a dog. Effective rewards can be either objective, such as verbal praise explicitly given to the model for approach behavior, or subjective and nonobservable, such as internal cognitions of social praise generated within the observing child (e.g., "That child is brave for approaching that dog. I will be brave for approaching that dog.") Although both tangible and cognitive stimuli may be equally effective in generating the initial performance of new approach behaviors, it seems likely that reinforcing cognitive self-statements (either those generated by the child spontaneously or those supplied by the therapist) are crucial in maintaining the new approach behavior. Similarly, the acquisition, as well as the therapeutic extinction, of phobic behavior may involve cognitive self-reinforcers, such as "She or he got knocked down by playing with that dog; I will get hurt if I play with that dog." Although the therapeutic effects of modeling with phobic children have been explored (e.g., Bandura, Grusec, & Menlove, 1967), little attention has been given to examining directly the nature and function of cognitive variables in the acquisition, maintenance, and extinction of children's phobias. In light of the role that such cognitions seem to play in phobic behavior, this is a major lack of the current research.

The two-factor theory incorporates both respondent and operant learning concepts. First proposed by Mowrer (1939), it postulates that phobias originate according to the respondent conditioning paradigm and are maintained according to the operant conditioning model. After initial respondent conditioning of fear or anxiety with the conditioned stimulus, anxiety reduction associated with avoidance of the noxious stimulus becomes a positive reward or reinforcer, since anxiety or fear is unpleasant. Thus, anxiety reduction becomes the reinforcement for avoiding the noxious stimulus created by respondent conditioning. Problems with the two-factor theory have been extensively discussed over the last several years (Bandura, 1969; Herrnstein, 1969; Rachman, 1976, 1977, 1978). In brief, (a) the theory assumes that phobias are mediated through the autonomic nervous system, although studies by Solomon and Turner (1962) and Bandura (1969) suggest that behavior is regulated in large part by the central nervous system; (b) it does not explain why people fail to acquire fears in what are theoretically fear-evoking situations (e.g., air raids); (c) it fails to explain the "choice of symptom" issue of why some stimuli are more likely than others to become fear signals, that is, the distribution of human fears is not consistent with the equipotentiality premise of the theory; (d) it accepts the faulty assumption that all fears are acquired directly through classical conditioning, although it has been shown that operant conditioning and observational learning can also indirectly produce fearful behavior; (e) it fails to explain why active avoidance behavior is so persistent and does not extinguish in the absence of further reconditioning or trauma experiences to maintain the anxiety associated with the conditioned fear stimulus. In response to some of these problems, Herrnstein (1969) has reported recent experiments showing that the conditioned stimulus may function as a *discriminative* stimulus for the avoidance response rather than as a noxious stimulus whose removal is inherently reinforcing, as two-factor theory requires.

Rejection of the two-factor theory has significant implications for both treating and measuring phobias, in that it suggests the three major components of fear (avoidance behavior, subjective experience, and physiological disturbance) can covary, vary inversely, or vary independently (Rachman & Hodgson, 1974). Routine tests of these three components should be carried out as a pretreatment measure to determine the amount of synchrony/desynchrony among the three elements and to determine which element(s) does (and does not) require direct treatment. Different treatments may tap different elements of a phobia more or less efficiently. Refutation of the two-factor theory also raises the important question of whether one needs

to treat each element directly or whether one can assure patients that if one cures one component (e.g., avoidance behavior), the others will follow automatically (e.g., subjective experience). Thus, although the two-factor theory itself does not yield any new treatment techniques, the theoretical and empirical debates over the theory have caused researchers to question both the necessity for and the sequencing of various therapy components based on respondent and operant paradigms.

The major theoretical models reviewed briefly here are not adequate to explain childhood phobias. Two issues that appear to us to be of major importance are not even addressed by the existing models. These issues are (a) the role of cognitions in the origin, maintenance, and reduction of fears and (b) the influence of developmental issues in childhood phobias. The idea that cognitions may influence phobic behavior is one that is gaining attention both on a theoretical level (Rachman, 1977) and on a treatment-oriented level aimed at fear reduction (e.g., Kanfer et al., 1975; Kissel, 1972; Meichenbaum & Turk, 1976). Developmental issues have long been ignored in this area, although recently some theorists (e.g., Bowlby, 1973; Kissel, 1972; L. C. Miller et al., 1974) have suggested that these issues may be important. Both cognitive and developmental issues deserve further exploration, which might lead to a more complete theoretical model of childhood fears and phobias.

There are two major categories of research in children's fears: normative fear survey research and fear-reduction studies.

Normative Fear Survey Research

It seems reasonable to expect that knowledge about the normal fears of children is important in understanding their pathological fears. At least 22 papers since 1932 have reported normative data, but they reveal little clear information. One point, however, is clear: The normative research has focused almost exclusively on the identification of fear stimuli by asking the question "What do children fear?" and then attempting to relate the number, type (content), or intensity of the fear stimuli to demographic parameters such as age, sex, and socioeconomic class and to other, presumably pathological, behavior. Most of the studies have focused on the number of fears, and some have focused on the content or type of fear, but very few have dealt with the intensity of fear. Despite many methodological problems, to be summarized later, some consistencies do emerge.

Sex Differences

One of the most consistent findings is that however fear is measured, girls obtain higher fear scores than boys (Angelino, Dollins, & Mech, 1956; Bamber, 1974; Croake, 1969; Croake & Knox, 1973; Cummings, 1944; Lapouse & Monk, 1959; Pratt, 1945; Russell, 1967; Scherer & Nakamura, 1968; Spiegler & Liebert, 1970). Although no study reported generally higher fears for boys, three articles (Maurer, 1965; L. C. Miller, Barrett, Hampe, & Noble, 1971; Nalven, 1970) found no sex differences.

Though the number of reported fears in relation to sex has been studied, there are uncertain data on the relationship between sex and either the content or the intensity of fears. Lapouse and Monk (1959) found significant sex differences in fear content, specifically in the percentage of children who feared certain objects. Pratt (1945), however, found no sex differences in fear content. MacFarlane, Allen, and Honzik (1954) collected data on fear intensity, but did not report any analysis. Other studies (Bamber, 1974; Russell, 1967; Scherer & Nakamura, 1968) found that girls report a greater fear intensity than boys. Although sex differences in content and intensity are still uncertain, the finding that girls report a greater number of fears appears frequently enough and with no conflicting evidence of reverse findings to be accepted as a reliable finding. Interpreting the finding, however, is difficult. One cannot tell from these data if the girls' higher scores reflect greater fear reactivity or if other factors, such as sex role expectations, operated. Consistent with a general role model of feminine behavior, girls may be more willing than boys to admit their fears. Similarly, in those studies using parents' reports, the adults may

incorrectly, but nevertheless reliably, attribute greater fear to girls than to boys. Grossberg and Wilson (1965) and Wilson (1967) suggested the operation of sex role factors in adult fear scores, whereas Scherer and Nakamura (1968) and Bamber (1974) suggested that similar sex role factors might account for the data on children and adolescents. One would expect, too, that if sex role stereotypes and behavior are changing, future fear surveys might find smaller and less reliable sex differences. In this regard it is interesting that the studies reporting no differences are fairly recent (from 1965 on), which possibly reflects some contemporary changes in sex stereotyping.

Age

Overall, there appears to be a general decrease from young childhood to late adolescence in the percentage of children who report one or more specific fears (Cummings, 1944, 1946; MacFarlane et al., 1954) and in the simple number of fears reported (Angelino & Shedd, 1953; MacFarlane et al., 1954; Nalven, 1970; Scherer & Nakamura, 1968). However, the decrease may not be in simple linear relationship with age, as several studies have shown a sharp increase in the number of reported fears around ages 9–11 (Angelino & Shedd, 1953; MacFarlane et al., 1954) or a peak at age 11 (Chazan, 1962; Morgan, 1959). On the other hand, several studies have reported no significant relationships of fear with age (Croake, 1969; Croake & Knox, 1973; Lapouse & Monk, 1959; Maurer, 1965; Russell, 1967).

Some studies suggest that the *type* of fear reported is related to age (Angelino & Shedd, 1953; Nalven, 1970; Scherer & Nakamura, 1968). Jersild and Holmes (1935a) observed that in infancy children's fears arise in response to occurrences in the immediate environment (e.g., loud noises, loss of support, etc.). As a child grows older, his or her range of fears grows wider and he or she acquires the ability to dwell on the past and to anticipate the future. Thus many of his or her fears will change to those of an anticipatory nature. This observation seems generally borne out by the data, as reviewed by Scherer and Nakamura (1968). The most consistent findings are an age-related decline in fear of animals (Angelino et al., 1956; Bauer, 1976; Lapouse & Monk, 1959; Maurer, 1965; Shepherd, Oppenheim, & Mitchell, 1972) and in fears of the dark or of imaginary creatures (Bauer, 1976; Holmes, 1936; Maurer, 1965; Shepherd et al., 1972) and an age-related increase in school and social fears (Angelino et al., 1956; Lapouse & Monk, 1959).

L. C. Miller et al. (1972b), in contrast with most other studies that showed an age-related decline in animal fears, reported that fear of small animals, along with "moral" fears, "emerges more clearly in adult life" (p. 268). L. C. Miller et al. extracted three factors as the "principle dimensions" of children's fear: (a) fear of physical injury, (b) fear of natural and supernatural dangers, and (c) fear of psychic stress (e.g., social events, examinations, etc.). The second factor was found to diminish with age, whereas the other two factors reportedly "emerge early and continue through much of the life span" (L. C. Miller et al., 1972b, p. 268). Similar factors have been reported by other investigators (Russell, 1967; Scherer & Nakamura, 1968).

An important developmental question concerns the relationship between age and degree or intensity of fear. Do children of different ages react differently to the same type and intensity of fear stimuli? It is commonly believed that young children have more intense or all-encompassing fear reactions than do older children. But to date there has been little investigation of this question.

Although detailed information on the relationship of age and the number, kind, intensity, and factor structure of children's fears is not known, there is consensus that age is an important variable in fear reactions.

In summary, as children grow older, their fear patterns change, but not in simple linear relation with age. Some fear stimuli remain operative, others lose their value, and some new ones emerge. These tentative conclusions about fear and age are drawn from a mass of research data. They appear reasonable but hardly surprising, and they seem a rather small output for so much research activity.

Socioeconomic Class

As with age, socioeconomic class (SEC) appears to be an important variable in children's fears, but the details of the relationships are not clear. Several studies have reported SEC differences in the type (content) or number of reported fears (Angelino et al., 1956; Bamber, 1974; Jersild & Holmes, 1935a; Nalven, 1970; Newstatter, 1938). The more reliable of the two findings is the variation found in content, with lower SEC children reporting more fears of specific events or others such as violence, whippings, dope peddlers, switchblades, drunks, money, rats, and cockroaches (Angelino et al., 1956; Nalven, 1970). In comparison, higher SEC children feared heights, car accidents, train wrecks, and large categories such as poisonous insects or dangerous animals. The fears of the lower SEC children ("ghetto" children in Nalven's, 1970, study) strongly suggest the socially determined nature of fear content and an immediacy and reality basis for the expressed fears of the lower SEC children. The data suggest that these children may perceive their immediate environments as far more hostile than do the higher SEC children. This is a hypothesis well worth testing, but it has not been investigated within the scope of the fear literature.

Although it seems clear that fear content varies with SEC, it is not clear whether the number of reported fears also varies. Some studies (Angelino et al., 1956; Bamber, 1974; Croake, 1969; Croake & Knox, 1973; Jersild, Markey, & Jersild, 1933) have found a greater number of fears for lower than for higher SEC children. However, as Garcia (Note 1) pointed out, Angelino et al. (1956), a frequently cited study, seemed to misinterpret their own data. The authors concluded that both content and number of fears varied with SEC, but their own data supported only the difference in content. Further, Nalven (1970) made an observation that may have bearing on this issue, namely, that lower SEC children compared with the others tended to list specific fear items rather than more generic groupings. For example, they listed specific animal fears (rats and cockroaches), whereas higher SEC children noted groupings such as dangerous animals or poisonous insects. If this difference in the level of abstraction of the concepts used by different SEC children was reliable, it may have generated a spurious finding of different numbers of fears between the groups. As Garcia suggested, studies that instruct children to simply list their fears are probably most susceptible to such an artifact. Finally, no studies have reported data on SEC differences in the intensity of fears.

Fears and Pathological Behavior

It seems a reasonable prediction that children's fears are positively related to other behaviors that are usually considered to be pathological. However, except for the suggestions of such a relationship found in clinical case studies, there seems to be little data to support this prediction. Some of the case studies (e.g., Tasto, 1969) have even pointed out that the child's severe fear was a specific and isolated reaction, apparently not related to other behaviors. Some studies, however, have shown a relationship between fear and manifest anxiety (Scherer & Nakamura, 1968) and between fear and neuroticism (Bamber, 1974). MacFarlane et al. (1954) reported that children's fears were related to a variety of problems such as general anxiety, irritability, and timidity, all depending on age and sex. Their correlations, however, were low and were not consistent across ages. Hampe et al. (1973) reported a general response to treatment over a 2-year follow-up period. As the primary fear reduced in time, so did "a host of other deviant behavior" (p. 451). These authors suggested that not only are fears related to other pathological behavior, but fears are related to other pathological behavior in some causative or mutually sustaining manner.

On the other hand, L. C. Miller et al. (1971) found no correlation between school phobia and other pathological behaviors, and Lapouse and Monk (1959), using detailed individual interviews with a random sample of the parents of 482 children, found no significant correlations between fears and pathological behaviors such as bed-wetting, nightmares, nail-biting, frequent temper losses, and so on. Lapouse and Monk also interviewed a sizable subsample of children

to compare their reports with the parents' reports and, again, found no significant relationships between fear and pathological behavior.

Thus there appears to be weak evidence that children's fears may be related to general conditions such as anxiety, but there is no clear support for the relation of fear with specific pathological behavior such as bed-wetting or tantrums. A reasonable hypothesis, not yet explored, is that the intensity rather than the simple number of fears may be positively related to behavior pathology. At this point, however, there is little evidence that childhood fears are related to other pathological behaviors.

Methodological Problems in the Normative Studies

For the most part methodologies have been fairly straightforward and simple, involving variations of questionnaire and interview methods that range from the earlier studies in which subjects were asked to list their fears to the more recent use of factor-analyzed rating scales. Only Jersild and Holmes (1935a) attempted to obtain normative data by direct observation methods.

In the earlier studies, children or parents were requested to "write down" or "tell me the things you are afraid of" (e.g., Maurer, 1965; Pratt, 1945) or to "list the fears of other children in your own age group" (e.g., Angelino et al., 1956; Nalven, 1970). Many studies using this list or list-and-rank method asked the children directly (Angelino et al., 1956; Croake, 1969; Maurer, 1965; Nalven, 1970; Pratt, 1945). Others, however, who sought greater reliability or perhaps more ease in obtaining data used mothers as informants (e.g., Hagman, 1932; Lapouse & Monk, 1959). Only Lapouse and Monk reported any measure of concurrent validity. They interviewed an additional sample of 193 children and their mothers separately, using different interviewers. Comparing those independent interviews, they found that the largest discrepancy concerned the mothers' and children's assessment of the number of children's fears and worries, with "the mothers, in comparison with the children, underestimating by 41 per cent" (Lapouse & Monk, 1959, p. 812). Their results cast doubt on the validity of mothers' or children's self-reports in this and other studies. Marks and Gelder (1966) asked adults to recall the childhood onset of their fears; Abe (1972), in a study carried out in Japan, compared self-report fears of 242 mothers with their childhood fears as recalled and reported by *their* mothers.

Of all the list-and-rank procedures, those of Lapouse and Monk (1959) appear to have been the most carefully executed. Using the Buffalo city directory, they drew a random urban sample and employed trained interviewers to conduct hour-and-a-half interviews with the mothers of 482 6–12-year-old children. The interviewers obtained data on the number and kinds of fears, on sex, age, race, and economic status, and on the occurrence of other behaviors such as bed-wetting, nightmares, and a variety of "tension phenomena" such as nail-biting. As mentioned, an additional sample of children and their mothers were also carefully interviewed to compare the children's and mothers' responses.

The list-and-rank approach is characteristic of the studies through 1966. The studies did identify fear stimuli, but they share a number of weaknesses. Two of the shortcomings specific to this approach are that one cannot be certain that the lists of fear stimuli are complete, and although data are generated on the number and kinds of fear stimuli, they tell us nothing about the degree of fear-stimulating power of each item or the intensity of the children's fear reactions.

Since 1967 some of the normative studies have been more sophisticated, but still have focused on the essential question, "*What* do children fear?" These later studies used rating scales that provided data not only on the number and types of fear stimuli but also on the relative degree of fear reactions to the different stimuli. In these studies, too, factor analyses were used to identify the presumed factor structure of children's fears.

The major approach in the rating scale studies is to instruct the child (Bamber, 1974; Russell, 1967; Scherer & Nakamura, 1968) or the parents (L. C. Miller, 1967; L. C. Miller et al., 1971; L. C. Miller et al.,

1972b) to rate each of the listed fear stimuli for intensity of reaction. The results are then factor analyzed, usually with added demographic or other behavioral measures.

Although fear-survey rating scales provide more information than the earlier approaches and have an advantage of standardization, their reliability and validity are still equivocal (Geer, 1965; Lang & Lazovik, 1963; Manosevitz & Lanyon, 1965). For example, L. C. Miller et al. (1971, 1972b) reported no validity data and only split-half reliability on the Louisville fear survey, perhaps the most often used fear rating scale for children. Although the split-half reliability was acceptable ($r = .80$), no data were reported on test–retest reliability or on any form of validity. Split-half reliability and factor analyses, the most commonly reported measures on fear rating scales, are measures of only the internal consistency of the scales. Additional data need to be collected on the stability (test–retest) and validity of fear rating scales before any but tentative conclusions can be drawn.

Taken together, the normative studies involve so many procedural differences and methodological problems that clear comparisons and reliable conclusions are difficult to draw. In addition to the difficulties specific to each method already pointed out, the field as a whole suffers from a number of generally shared problems. One major limitation is that the data are essentially subjective, that is, they are usually second-person reports by parents and sometimes first-person reports by children. Further, the adult second-person reports are usually limited to mothers of the children, with whatever sex role biases and attributions may be operative. We do not know, for example, if fathers' reports would generate different results. Perhaps it might be hypothesized that the sex differences would appear even greater, with fathers attributing even less fear to their sons than to their daughters. Whether made by parents or by children, the reports are based on unsystematic observations that are subject to many biases, and the respondents' recall is subject to many distortions. Neither test–retest reliability nor validity measures are routinely taken.

Although normative studies have been reported since 1932, their results are not generalizable because random sampling from the normal child population is seldom attempted. With very few exceptions (e.g., Lapouse & Monk, 1959; L. C. Miller et al., 1971), these are not properly normative studies. Investigators appear to have been interested in limited segments of the population or were restricted in terms of the availability of subjects. For example, some studies (Marks & Gelder, 1966; Poznanski, 1973) used samples drawn from psychiatric populations, and Maurer (1965) used children referred to school psychologists. Abe (1972) obtained data from a self-selected population of young mothers who were patients in a free medical clinic in Japan; Bamber (1974) studied children in Ireland; and Pratt (1945) studied only rural children.

In summary, the basic subjective-report methods have obvious shortcomings: some studies interviewed children, some parents; some required lists, some ratings; the ages of children are not comparable across studies; random samplings from normal childhood populations were seldom taken; reliability and validity data were not reported.

But the major limitation in all of this research is the nature of the questions that have typically been asked. The studies have been virtually limited to identifying fear stimuli, to asking, essentially, "*What* are the common fears of children?" and "To what demographic factors do they relate?" Most studies seem still to be trying to accomplish Hagman's (1932) first aim, that is, "to enumerate and analyze the objects and situations feared" (p. 110). It seems clear that many factors influence the objects or situations chosen as fear stimuli. Some stimuli, as Maurer (1965) pointed out, seem to be merely the subject of the latest adventure story or movie to which the child was exposed; others seem to be stereotypes or culturally conditioned fears, and still others appear to have some significance for the child's life (e.g., Angelino et al., 1956; Maurer, 1965). The type of identification and tracking of fear stimuli that has been done could probably be continued indefinitely. However, the question must be raised as to the point

of such an enterprise. In an area as broad and important as fear in children one must ask oneself what kind of research is worth doing and what kind of knowledge is worth having.

The identification of fear stimuli touches on only a small part of what must be a complex fear process. It seems reasonable to assume that children's fears, like other human reactions, proceed through complex paradigms—from fear stimuli that vary in number, type, and intensity and that may be internal, external, or both; through emotional and cognitive operations within the child; through overt fear responses that may act upon and modify both the social and physical environment and that themselves, by means of feedback loops and chaining, may occasion variations in any part of the process. The processes themselves may further vary with developmental factors. Given this reasonable assumption of human complexity, it is clear that the normative research to date has ignored a great deal of it, limiting its view to only one part of the process, that is, identifying fear stimuli. It is as if early researchers took the first reasonable steps toward learning what it is that children typically fear, but then allowed the field to remain on that level, the researchers generating, sharing, and rearranging virtually the same lists of fear stimuli. The research tells us little about the remainder of the paradigm, the nature of children's fear experiences. Remaining unasked are questions concerning the operation of mediating cognitions in children's fear experiences, the degree to which fears are self- or externally generated and maintained, and the effects of fear behavior on the child's social environment and the effects' feedback influence on the child. One must investigate how children in their natural environments typically deal with fearful events and how their strategies vary with developmental level, sex, and so on. One must study the conditions under which natural coping processes fail and fear processes become debilitating, and one must determine the optimum conditions for fear-reduction intervention.

We suggest, too, that there is a related, potentially important issue that has thus far been ignored in the literature: the recognition of the possible adaptive value of childhood fear. Of course one has no clear indication that fear serves any adaptive function at all. But it may be that age-linked, transitory fears have important short-term effects on coping with the social environment, on learning how to appropriately sensitize and desensitize oneself. Fear may constitute an important part of children's experiences in learning to successfully cope with problems. The possible adaptive value of fear experiences in children's development may be an important issue, but has been overlooked, clouded perhaps by an implicit, unquestioned value judgment that fears are only disruptive and hence are of negative value.

A great variety of stimuli have been identified as having fear-stimulus value; they range from specific objects (e.g., bees) to abstract and imaginative stimuli such as communism and ghosts. As pointed out by Berecz (1968) and L. C. Miller et al. (1974), almost any event may be a potential fear stimulus. Thus children have access to an inexhaustible supply of them. The normative research suggests that "normal" fear stimuli for children are socially determined and are appropriate for the individual's personal and social situation; for example, the elderly fear loneliness and physical injury, students fear examinations, lower SEC boys fear switchblades and beatings, and young children, still learning the limits of reality, fear ghosts, witches, and darkness. The fears are appropriate to age, social class and role, culture, and even moment in history. Thus the normative data suggest that what is feared by children is largely determined by social and historic fashion as well as by individual experiences. Although some fears may be innately determined, it appears that children fear largely what they are taught to fear. It seems that identification of fear stimuli may be limited in its contribution to understanding the processes of learned fears in children, since so many stimuli may be interchangeable. Further research in this direction is of doubtful value, and attention to other parts of the complex fear paradigm would be more profitable. Overall, the normative research only suggests, and with a good amount of conflicting data, that the type

and number of fears vary with age, sex, and SEC. The details of the apparent relationships, however, are not clear. There is conflicting evidence concerning the relationship of fears to pathological behaviors, and there are virtually no data on the intensity of fear reactions. The normative research reveals very little beyond the findings of what appear to be a few key studies (e.g., Jersild & Holmes, 1935a, 1935b; Lapouse & Monk, 1959; L. C. Miller, Barrett, Hampe, & Noble, 1972a; L. C. Miller et al., 1972b). The "disappointing research" (Berecz, 1968) may be at least partly due to the ethical problems of fear research with children as well as to a variety of methodological problems that characterize the research. Because of its apparent fixed focus on identifying fear stimuli, the research tells one little about the complete fear processes.

Case Studies of Fear Reduction

Like other aspects of the literature in this area, fear-reduction experimental and case studies have a long history. According to Yates (1970), most of the techniques that are now labeled *behavior therapy* were applied in one form or another some 50 years ago, almost entirely in attempts to treat children's fears. Watson and Rayner (1920) suggested, but did not investigate, four possible methods to remove the fear that they had conditioned in Little Albert. Jones's (1924b) treatment of Peter is an early example of the application of learning principles to the remediation of clinical problems. Jones (1924a) studied 70 children in an attempt to discover ways in which children's fears could be reduced or removed. Holmes (1936), stressing the importance of active participation by the child, attempted to eliminate fears of the dark and of heights, whereas Hagman (1932) and Jersild and Holmes (1935b) reviewed methods that parents actually used to deal with fears and suggested a number of more effective techniques.

Unfortunately, as with the normative studies, the promise of the early work was not fulfilled. The vigorous interest in treatment of children's fears seemed to disappear, and except for a few case studies, there was little further activity until Rachman and Costello's (1961) review of the etiology and treatment of children's phobias. They called for "above all a major project to establish the degree and permanence of improvement which may be obtained by these techniques" (p. 104) (i.e., those suggested by Jones, 1924a, 1924b, and by Jersild & Holmes, 1935b). This call has also gone largely unanswered. The amount of research in the behavioral treatment of children's fears since that time has been surprisingly meager in comparison with the amount of behavioral research on adult fears, an area with well over 100 outcome studies on systematic desensitization alone (Wolpe, Brady, Serber, Agras, & Liberman, 1973).

Most of the reports on childhood fears are single-case studies, with all of their attendant limitations. However, case studies may offer something that experimental studies do not, that is, examinations of children's phobias at high levels of severity, and may give hints about the nature of phobias and treatment suggestions.

Prior to 1960 there were only a few behaviorally oriented case studies of child fear reduction (e.g., Jones, 1924b; Rodriguez, Rodriguez, & Eisenberg, 1959; Weber, 1936). Since 1960 at least 35 additional papers have appeared, which report on 112 children referred for clinical treatment. Of these papers, 20 were limited to school phobias and account for 72 of the cases. The remaining 15 case studies, accounting for 40 children, involved school phobias plus a variety of other fears. Thus, since 1960, excluding school phobias, all other children's fears averaged about 2 case studies per year, a strikingly low average compared with the rest of the behavior therapy field.

Considering the variety of fears reported in normative studies and included in fear surveys, the range of phobias that are presented for treatment is quite small. The overwhelming majority of reported cases involve school phobia, followed by those involving fears of animals, noise, and bodily injury. Even assuming that most of these clinical cases represent severe levels of fear, we note that both the range and the number of severe fears presented for treatment are

relatively small. Thus, the important question remains unanswered: Are severe fears largely undetected, or do they simply not occur at a very high rate?

The school phobia literature has been reviewed (Gelfand, 1978; Hersen, 1971; Kelly, 1973), and this section is presented only as a brief overview. School phobia is more prevalent than other childhood phobias (L. C. Miller et al., 1974), a fact that may be due to the high cultural value and legal requirements placed on school attendance. Because it is socially and legally less acceptable for a child to avoid school than to avoid dark places, small animals, and so on, more school phobia cases are likely to be referred for clinical help, that is, the school-phobic child conflicts with social convention more visibly and inconveniently than other fearful children.

Previous reviews suggest that the prognosis for improvement is good with most types of treatment, provided that school refusal has not yet become a chronic pattern. Behavior therapy procedures for school phobia generally are based on both classical and operant conditioning models. Thus, school phobia is seen as avoidance behavior motivated by high anxiety and maintained by reinforcement for not attending school. Four major respondent-based procedures have been reported: (a) systematic desensitization (Chapel, 1967; Lazarus, 1960; Lazarus & Abramovitz, 1962; P. M. Miller, 1972); (b) in vivo desensitization (Garvey & Hegrenes, 1966; P. M. Miller, 1972; Olsen & Coleman, 1967; Tahmisian & McReynolds, 1971); (c) flooding (Kennedy, 1965); and (d) implosion (Smith & Sharpe, 1970). Three major operant conditioning procedures were reported: (a) home-based contingency management (Ayllon, Smith, & Rogers, 1970; Cooper, 1973; Edlund, 1971; Hersen, 1970; Kennedy, 1965; Vaal, 1973); (b) school-based contingency management (Brown, Copeland, & Hall, 1974; Hersen, 1970; Rhines, 1973; Weinberger, Leventhal, & Beckman, 1973), and (c) behavioral shaping in the clinic (Hersen, 1970; Patterson, 1965). Most case studies report procedures based on combinations of the classical and operant models, with emphasis on one. The choice of which theoretical model and therapeutic procedure to use seems determined by whether the most pressing therapeutic need is to reduce the child's anxiety with a desensitizing procedure or to avoid reinforcing his or her escape behavior. Follow-up data (ranging from 4 weeks to 2 years) were reported for the majority of cases, although the data were limited either to assessment of continued school attendance or to attendance data plus a verbal report from parents and teachers on the child's social or emotional functioning. Actual behavioral assessment of improvement after return to school was not reported for any case. Whether such follow-up data should be collected is apparently still debatable. Hersen (1971) has criticized behaviorally oriented therapists for not being more rigorous in obtaining data regarding academic, social, and emotional adjustment. However, Ayllon et al. (1970) stated that "school attendance is a legitimate if not the only relevant treatment objective for school phobia" (p. 135).

Behavior therapy procedures for phobias other than school phobias are generally based on a classical conditioning model, that is, the avoidance is seen to be mediated by high anxiety. Two main approaches are used to reduce the inferred mediating anxiety: (a) counterconditioning procedures that pair the anxiety response with a stronger response that is also antagonistic to anxiety, such as muscular relaxation, and (b) extinction procedures in which the anxiety response is repeatedly elicited, but presumably without reinforcement. Therapy for these phobias usually involves a combination of several procedures. The terminology used for the procedures is confusing and overlapping, but the various authors claim that real differences do exist, as summarized by the following four categories: (a) reciprocal inhibition (Bentler, 1962; Freeman, Roy, & Hemmick, 1976; Jones, 1924a; Katz, 1974; Lazarus & Abramovitz, 1962; Miklich, 1973; Ney, 1968; Tasto, 1969; Wish, Hasazi, & Jurgela, 1973), (b) contact desensitization (Weber, 1936), (c) implosion/flooding (Ollendick & Gruen, 1972; Smith & Sharpe, 1970), and (d) in vivo desensitization plus guided practice (Croghan & Musante, 1975). In cases using reciprocal inhibition, the following responses antagonistic to anxiety were used: feeding (Jones, 1924a), muscle relaxation (Tasto, 1969),

playing (Bentler, 1962; Ney, 1968), and emotive imagery (Lazarus & Abramovitz, 1962). Interestingly, playing was included as an integral part of several treatment procedures (Bentler, 1962; Croghan & Musante, 1975; Ney, 1968; Weber, 1936). However, none of these authors stated explicitly that playing was used to teach the child skills of interaction with the feared stimulus. This lack of emphasis on skills training in case studies is interesting in light of Rimm and Masters's (1974) suggestion that therapists not place sole reliance on decreasing anxiety through desensitization techniques when working with socially inexperienced young children, who often lack the appropriate behavioral skills required to cope with the feared stimulus. Follow-up data were reported in 10 of the 15 recent papers. However, such data were rarely behavioral, making comparisons among cases regarding long-term effectiveness difficult, if not impossible. Rather, follow-up data were obtained in verbal reports of parents or teachers regarding the child's maintenance of his or her new approach behavior.

In summary, we can draw the following tentative conclusions about the clinical treatment of childhood phobias, as reported in case studies. First, given the strikingly low number and narrow range of reported cases, it seems that childhood phobias may not be a significantly large clinical problem, except for school phobia. Second, excluding school phobia, the most frequently used therapy procedures are modifications of those used with adults, that is, systematic desensitization and implosion/flooding. As has been noted elsewhere (Graziano, 1975), serious ethical and humanitarian questions are raised in using implosion/flooding techniques on children, because children generally have no choice in whether to enter and remain in therapy. Furthermore, as Ullmann and Krasner (1975) have cautioned, the successful use of implosion techniques requires considerable clinical skill and sensitivity so that the highly aversive treatment experience is associated with the avoided object and not with the therapist. In contrast with other phobias, school phobia is again an exception in that its treatment procedures generally include the use of both classical and operant techniques. Thus, this set of case studies may be taken as a comment on the state of the field: With the exception of school phobia, the clinical importance and prevalence of childhood phobias is questionable, and the procedures used are generally copied from work with adult phobics.

Important therapy and research questions raised by these case studies appear to fall into three main categories. First, what procedures not generally used with adults might be useful with children? As noted by Rimm and Masters (1974), young children often lack the appropriate behavioral skills required to cope with the feared stimulus. Thus, the absence of much if any overt skill training in the case reports seems a glaring omission. Such skill training might take the form of guided participation or modeling. Second, what treatment combinations are most effective and efficient for different kinds of phobic children? As pointed out by both Berecz (1968) and Hersen (1971), more controlled research designs are needed to adequately test both the efficacy of any one procedure and the conditions under which it will be most effective, both alone and in combination with other procedures. Case studies are limited in two senses: (a) One has no way to know whether other attempts to treat a phobic child by these authors or other authors using similar techniques were unsuccessful (i.e., how unique is the success of this procedure?); (b) we cannot compare the effectiveness of two or more different procedures because most reports use language that cannot be interpreted in terms of specific behavioral gains (e.g., "has remained in school and continued to improve"). Thus, despite the difficulties, a clear need still exists for more controlled designs using experimental groups or highly systematic multiple studies. Finally, these case studies raise the question of what criteria for success should be used, both at treatment termination and at follow-up. Should one assess only improvement in approach behavior? Or should one also assess improvement in social and emotional functioning? One rationale for collecting social and emotional adjustment data is that these data might help researchers make comparisons among the various techniques, with the

ultimate aim of identifying the most rapid and widely effective combinations. At any rate, termination and follow-up data should be collected and reported in terms of behavioral observations as well as in terms of verbal reports from parents and teachers.

Controlled Fear-Reduction Studies

At least 28 controlled fear-reduction studies with children have been published, and there are several additional unpublished papers such as those reported in Gelfand's (1978) review. Although far more rigorous and systematic than the clinical case studies, the controlled studies share several common limitations. First, with few exceptions (e.g., L. C. Miller et al., 1972a; Obler & Terwilliger, 1970), they study what appear to be mild to moderate levels of fear, and the relevance of their findings for treating children with severe fears or phobias has yet to be demonstrated. Second, as with the case studies, the range of fears included has been limited: 9 studies focused on dental/medical fears (Adelson & Goldfried, 1970; Adelson, Liebert, Poulos, & Herskovitz, 1972; Johnson & Machen, 1973; Machen & Johnson, 1974; Melamed, Hawes, Heiby, & Glick, 1975; Melamed & Siegel, 1975; Melamed, Weinstein, Hawes, & Katin-Borland, 1975; Vernon & Bailey, 1974; White & Davis, 1974). Animal fears were discussed in 7 studies (Bandura et al., 1967; Bandura & Menlove, 1968; Hill, Liebert, & Mott, 1968; Kornhaber & Schroeder, 1975; Murphy & Bootzin, 1973; Obler & Terwilliger, 1970; Ritter, 1968). Fears of social interaction were discussed in 5 studies (Evers & Schwarz, 1973; Keller & Carlson, 1974; O'Connor, 1969; O'Connor, 1972; Walker & Hops, 1973). Three studies focused on fears of the dark (Kanfer et al., 1975; Kelley, 1976; Leitenberg & Callahan, 1973). Two discussed test anxiety (Mann, 1972; Mann & Rosenthal, 1969), and 1 studied fear of water (Lewis, 1974). L. C. Miller et al.'s (1972a) study included a variety of fears, such as fear of school, the dark, heights, germs, nakedness, and storms. However, if one deletes school phobics, the number of children in L. C. Miller et al. who presented fears other than those typically studied totals only 11.

Another general limitation is that most studies use a number of different techniques in complex treatment packages, and it is usually not possible to determine the most effective components of the packages.

The profusion of concepts and methods makes it difficult to organize these studies for discussion. They present a variety of terms, including desensitization, extinction, modeling, reciprocal inhibition, and so on. It is frequently unclear whether a term, systematic desensitization, for example, is used to denote a treatment procedure and is thus part of the independent variable or whether it refers to implicit processes of fear reduction within the child and is thus part of the dependent variable. Likewise, apparently similar processes or procedures are labeled differently by different authors, such as Ritter's (1968) "contact desensitization" versus Lewis's (1974) "participant modeling."

Most authors make their procedures fairly explicit, but only imply their models of fear-reduction processes. However, it appears that all of these authors assumed a disinhibition model, that is, the operation of some process of gradual weakening of the fear response in the course of graduated exposure to the fear stimuli. The graduated exposure may occur actually, imaginally, or vicariously.

The methods used to operationalize gradual exposure fall into three main groups: (a) modeling, (b) multivariable, systematic desensitization or contingency management treatment packages, and (c) a cognitive approach using verbal-mediation strategies. By far, modeling is the most frequently used procedure, accounting for 20 of the 28 controlled studies. Many of the remaining 8 papers seem to have included some forms of modeling, but did not label them as such. Modeling may be implicit in any of the fear-reduction procedures except, perhaps, for imaginal systematic desensitization.

Modeling

Modeling is largely a development of the last decade; it was stimulated by Bandura's research (e.g., Bandura et al., 1967; Bandura & Menlove, 1968), although precursors of modeling are clearly seen in Jones (1924a).

At least 20 modeling studies of fear reduction in children have been published in the past 10 years. Although there are clear similarities among them, they fall into two fairly distinct groups with a number of important differences. Of these studies, 11, including Bandura's research, dealt with fairly common fears that occur frequently, even daily, for many children in the course of everyday events (i.e., fear of animals, the dark, water in baths, school examinations, and social interactions). In most instances the fears are clearly unreasonable, in that there seems to be nothing objectively harmful in the immediate situation. The other 9 studies form a distinct group of recent research published in the 1970s. They focused on the specific problems of reducing children's fears and preparing them for dental and surgical procedures. Unlike those in the former group, these fears are highly situation-specific and include often intense, but transitory, anxiety; they occur only at those few times when children are faced with dental or surgical procedures and thus, unlike many other fears, are not frequent or everyday issues. Also, the fears in this group appear more reasonable or reality based than, for example, fears of the dark. It does not appear to us to be at all unreasonable for a child to struggle against a dentist or surgeon who is going to do unexplained things that, as far as the child knows, will probably hurt a great deal. In fact, we would argue that such self-defensive assertion on the part of a frightened child who is psychologically unprepared for the coming medical treatment is preferable to quiet, but totally frightened, submission. This group of modeling studies focused on the reduction of such fears in these specific situations. They may thus be of immense potential importance for untold numbers of children.

One of the studies in this group (see the listing above) is Melamed and Siegel's (1975) excellent research using symbolic modeling procedures to prepare children psychologically for surgery. The experimenters used a 16-minute film of an initially fearful child who gradually copes with the situation and overcomes his or her own fears. They reported that the children who viewed the modeling film, compared with control group children, showed significantly less "transitory, situational anxiety" on all measures (sweat gland activity, self-reported medical concerns, and overt, anxiety-related behavior). Their data also suggested that experimental group children had fewer postoperative behavior problems than did the control group children.

These 9 studies give good evidence that brief symbolic modeling may be an effective aid in preparing children for dental and medical treatment, thus reducing their possibly high situational fears. The studies are limited to some degree in that they studied groups of normal children and did not include dental or medical phobics per se, and thus one cannot generalize these results to such chronically fearful children. Further, there is little evidence of the long-term effects of such modeling and yet no differential study of the effect of variables such as cognitive rehearsal, model and subject similarity, age, and so on. But all of those limitations can be avoided in future research.

In all of these studies the strategy was a preventive one; that is, prior to the occurrence of the stressful situation, children were prepared so they could more effectively cope with the stress when it did occur. There are important implications here for preventive and "stress inoculation" approaches, as is briefly discussed later.

Overall, the quality of these 20 modeling studies is high, but generalization are limited for a number of reasons. First, none of the studies typically included severely fearful children, although Bandura and Menlove (1968) did describe fairly high levels of dog fear in some children, and many of the subjects in the dental/medical fear-reduction research seemed quite fearful. Overall, the children were not selected for severe or phobic levels of intensity or duration of fears. Although modeling appears to be quite effective with mild to moderate fears, its usefulness in reducing phobias or severe fears has not yet been shown. In order for modeling to be effective, the child must attend to the model. Children who are highly fearful may find watching any interaction with the phobic stimulus so aversive that they look away, thus avoiding the fear stimulus. Bandura and Menlove (1968) commented on encoun-

tering a similar difficulty. Thus, successful modeling with severely fearful or phobic children has yet to be demonstrated.

All of the studies used behavioral avoidance tests or behavioral observations in their pretreatment assessments. These procedures separate children by fear levels in comparison with other children in a particular study, but do not assess the clinical significance, per se, of the fears. Further, although the subjects' age range was wide (3–13 years), selective factors may have operated, for example, the use of children from special university nursery schools (e.g., the Bandura studies) or from parochial schools (e.g., Kornhaber & Schroeder, 1975) or the use of only black children (e.g., Lewis, 1974) or only poor center city children (Melamed, Hawes, Heiby, & Glick, 1975). Thus, the subjects in the experimental studies of modeling are not representative of either a clinical population or the normal population of children.

As noted earlier, the range of fears to which modeling has been applied is limited, with nine studies focusing on fears of impending dental/medical treatment, six on animal fears, four on social isolation, and one on fear of water. Thus one must be cautious about generalizing from these modeling studies either to the general or clinical populations or to different fear stimuli.

The basic modeling approach common to these studies consists of fearful subjects' first observing models who demonstrate approach to the fearful stimuli. The subjects then attempt to perform the approach behavior themselves. This basic method has been varied across several dimensions: The modeling has been symbolic, as on videotapes (e.g., Bandura & Menlove, 1968; O'Connor, 1972), or presented by live models (Bandura et al., 1967; White & Davis, 1974); subjects have observed a single model (Ritter, 1968) or multiple models (Bandura & Menlove, 1968); modeling alone has been compared with modeling plus active, graduated contact with models and the fear stimulus (Lewis, 1974; Murphy & Bootzin, 1973; Ritter, 1968); the models have approached single stimuli (e.g., Kornhaber & Schroeder, 1975) or graduated variations of the fear stimulus, such as Bandura and Menlove's (1968) use of a variety of dogs; and the similarity of models to subjects has been explored (Kornhaber & Schroeder, 1975).

Not all of the meaningful combinations of these and other variables have yet been examined, nor have sufficient replications of successful studies been carried out. However, some clear consistencies and at least tentative conclusions have emerged. First, therapeutic modeling procedures are in general significantly more effective than control conditions and can override pretreatment differences in "predisposition to emotionality" (Bandura & Menlove, 1968) and in model's similarity to the subject (Kornhaber & Schroeder, 1975). Further, modeling procedures seem effective over a fairly wide age range.

Second, the research suggests that modeling procedures become more powerful as additional controlled components are added to the basic techniques. Thus it seems likely that a highly effective modeling package would include multiple live models who approach a hierarchical range of varied fear stimuli, followed by trials of active contact with models in graduated sequences of progressively bolder contact with the graduated fear stimuli (Murphy & Bootzin, 1973; Ritter, 1968), that is, a modeling plus contact-desensitization package. There is considerable agreement that live modeling combined with progressive contact with models and fear stimuli is potentially a very powerful technique. To date, however, the complete modeling package outlined here has not been tested either against control or against other modeling procedures.

Third, although live modeling appears to be more effective than symbolic modeling, the latter may be equally effective (Bandura & Menlove, 1968) if it involves many trials and includes multiple models and progressively varied fear stimuli. In other words, a symbolic modeling package similar to that described above for live modeling may also be a powerful fear-reduction technique. The significance of this lies in the potential use of symbolic modeling procedures in large-scale fear-reduction projects, in which direct contact with large numbers of children is not feasible. Thus, to the degree that children's fears may be a large-scale problem, symbolic

modeling may offer potentially useful remediation. As of now, the effectiveness of such a program presented, for example, in educational television broadcasts, has not been tested and remains only a possibility.

Fourth, symbolic modeling has important implications for immunological approaches (e.g., stress inoculation; Meichenbaum & Turk, 1976). Although Bandura and Menlove (1968) pointed out the possibility of such preventive strategies as early as 1968, only recently have such approaches been applied to children's fears. Poser and King (1975) reported on a series of their experiments on the reduction of avoidance. In every case, the avoidance inhibitor was brought to bear before rather than after avoidance responding had occurred. One experiment made use of symbolic modeling to inhibit the acquisition of snake avoidance responses in first graders confronted with a live snake for the first time. These approaches deserve much more attention. For example, as is suggested by studies such as O'Connor's (1972), might it not be possible to "immunize" entering kindergarten or nursey school children against early school fears? The symbolic modeling studies on reducing dental/medical fears also have clear implications for fear prevention. Large-scale televised symbolic modeling is a potential preventive strategy that has important community mental health implications. It is an area well worth careful attention.

Fifth, some dimensions of model and subject similarity may be important. Kornhaber and Schroeder (1975) found that age similarity (peer modeling) was associated with greater fear reduction for children than was adult modeling, although neither attitudinal nor behavioral similarity was an effective variable. Although researchers have investigated model–subject similarity with regard to other behaviors (e.g., Bandura & Barab, 1973; Bandura, Ross, & Ross, 1973; Baron, 1971; Hicks, 1965; Maccoby & Wilson, 1957; Wheeler & Levine, 1967), only Kornhaber and Schroeder (1975) have manipulated model similarity in child fear reduction. Clearly this issue—the relationship between model and subject similarity and the reduction of children's fears—is an area in which more research is needed.

Sixth, Kornhaber and Schroeder (1975) and Melamed and Siegel (1975) also touched upon a related and possibly important issue, that is, the relative effectiveness of models' portraying a *mastery* versus a *coping* strategy in fear reduction. Several authors (Blanchard, 1970; Rachman, 1972; Ross, 1970) have suggested that modeling effectiveness might be increased if models demonstrated initially fearful behavior that gradually changed to fearless interaction with the stimulus. Hill et al. (1968), Melamed and Siegel (1975), Meichenbaum (1971), and Spiegler, Liebert, McMains, and Fernandez (1969) all had their models exhibit some fear during the initial approach. Using adult subjects, Meichenbaum (1971) directly compared coping (initial display of fear) and mastery (no fear modeled) strategies. Meichenbaum found greater fear-reduction effectiveness with the coping strategy than with the mastery strategy. However, in all three studies there was sufficient confounding of variables to make their results uncertain. In the Meichenbaum study, for example, the coping models—but not the mastery models—also demonstrated a breath-control relaxation technique that may have accounted for many of the differences.

Although Kornhaber and Schroeder (1975) compared a coping and a mastery modeling procedure, their coping models displayed fear throughout the modeling and did not demonstrate a progressive reduction of fearful behavior that matched the demonstration of increased approach. Their coping model was thus incomplete. Finally, the coping model procedure used by Melamed, Hawes, Heiby, and Glick (1975) and Melamed and Siegel (1975) was part of the modeling package applied to all experimental subjects and was not compared with a mastery model condition, thus making it impossible to assess the effects that can be attributed to coping strategies per se. To date no researchers have systematically investigated and carefully compared coping versus mastery modeling procedures in child fear reduction. This is an obvious area in need of research.

Seventh, although modeling procedures appear to be of a decidedly social nature, they seem to have rarely been used to reduce basically social fears of children such as

shyness, severe isolation, fear of speaking in groups, and so on. An exception is O'Connor's (1969, 1972) work with socially withdrawn nursery school children. Using only a 23-minute peer-modeling film that showed progressively more involved social interactions, O'Connor reported dramatic changes in originally shy nursery school children. The children reportedly became normally assertive in class and "indistinguishable" from their nonshy classmates after only one viewing. Control children did not change. O'Connor's results are extremely impressive, but, as discussed by Gelfand (1978), a number of subsequent attempts have failed to demonstrate positive effects of symbolic modeling on social withdrawal. The experimenters, nearly all in unpublished papers, attempted to explain their lack of success by pointing to their own procedural shortcomings (e.g., too disparate model-subject ages, lack of a sound track to accompany a modeling film, a failure to clearly model coping strategies, etc.). These failures suggest that modeling may not be as powerful and reliable a technique as in suggested by the published studies. It is most interesting that the attempts to account for the failures suggest a possibly strong presumption of the effectiveness of modeling strategies, a presumption powerful enough to necessitate explaining away the failures as experimental weaknesses. In other words, there may be a tendency to discount negative evidence on modeling.

One of the successful modeling studies on social withdrawal (Jakibchuk & Smeriglio, 1976) reported that the videotaped model's use of a first-person narrative in which the child described himself or herself as initially shy was essential. Their work suggests the operation of important cognitive factors through first-person self-statements as well as the importance of the model's behavioral similarity to the observing child.

Finally, even when modeling is successful in modifying fear behavior, the effects of modeling may disappear in a few weeks for adults (e.g., Bandura, Jeffery, & Wright, 1974) and for children (Keller & Carlson, 1974) if subsequent reinforced practice is not made available to the observer. Like other behavior modification techniques, particularly those used in the 1960s (Graziano, 1975), modeling with fearful children has repeatedly demonstrated its success in modifying target behavior; but there are yet few data on either the generalization or the maintenance of the new behavior. Further research on these two issues is necessary.

Overall, despite some negative evidence and many yet unanswered questions, the bulk of the research evidence supports modeling as an effective set of techniques and a rich area for further research. Needed are systematic examinations of the effects of several variables (e.g., model and subject similarity, coping and mastery strategies, single and multiple modeling, etc.). Three of the more important questions to be answered concern (a) the intensity of fears for which modeling might be effective, (b) the potential use of large-scale symbolic modeling to reduce and, perhaps of greater importance, to prevent particularly common and disruptive fears, and (c) the generalizability and maintenance of behavior newly achieved through modeling. Conspicuously absent in the modeling research to date is direct investigation of the degree to which cognitive factors such as covert competency self-statements are operative.

Desensitization and Contingency Management Packages

Seven studies describing complex treatment packages are included here. Two studies (L. C. Miller et al., 1972a; Obler & Terwilliger, 1970) included children with severe fears. Two other studies (Mann, 1972; Mann & Rosenthal, 1969) studied adolescents referred by school counselors for test anxiety who could also be considered at or near severe levels. Obler and Terwilliger used systematic desensitization to pictures of fear stimuli followed by up to 50 hours of in vivo contact desensitization to reduce the severe monophobias (of dogs or buses) of 30 "neurologically impaired" children.

L. C. Miller et al. (1972a) compared reciprocal inhibition, psychotherapy, and a waiting-list control condition as treatments in phobic children. The reciprocal inhibition treatment condition was actually a broad array of behavioral approaches, including

meetings with parents, restructuring home contingencies, assertive training, relaxation and systematic desensitization, and discussions of problems and progress. The psychotherapy treatments focused on inner experience and encouraged the child to talk about his or her feelings and conflicts. There seems to have been considerable overlap in these two treatment conditions: For example, discussions with parents and altering home conditions are similar strategies; and denying the child in the reciprocal inhibition group free access to television at home ("restructuring contingency schedules") to decrease the reinforcement for fear behaviors is similar to the strategy of "removing gratifications" so as to reduce "secondary gains."

Their results are complex and difficult to summarize. Overall, neither treatment was more effective than the waiting-list control treatment. However, when age is considered, the reciprocal inhibition and psychotherapy treatment conditions were equally more effective than the control conditions for children 10 years of age and younger, whereas those 11–15 years old did not seem to respond to either treatment condition. In a 2-year follow-up (Hampe et al., 1973) the investigators found major effects of age and elapsed time. The younger children had "changed dramatically" immediately following treatment and maintained their improvement. Older children showed a more gradual improvement over the 2-year follow-up, as did control children. The researchers concluded that age and time are critical factors in child fear reduction. Younger children tend to lose their fears much more rapidly than do older children; in time (up to 2 years), children generally lose their fears with or without treatment; nevertheless, treatment, behavioral or psychodynamic, is justified because it "greatly hastens recovery" (p. 451).

Both studies, particularly Obler and Terwilliger's, employed complex treatment packages. Although both research groups labeled their major procedures *systematic desensitization* (Obler & Terwilliger, 1970) or reciprocal inhibition (L. C. Miller et al., 1972a), that is, some variation of Wolpe's (1974) systematic desensitization, they both also included a variety of contingency management techniques, including shaping and probably some modeling. Neither study clearly specified the roles of the therapists, and it is difficult to know just what occurred in the various treatment groups. Further, both studies used parents' reports as the major dependent variable. In fact, Obler and Terwilliger based their analysis on a single item from a 10-item parents' questionnaire. As critiqued by Begelman and Hersen (1971), Obler and Terwilliger's study has methodological problems that make its positive findings highly questionable. L. C. Miller et al. (1972a) and Hampe et al. (1973) are useful studies, in spite of the difficulties noted above. They present important findings concerning fear reduction as a function of age and time, and they include a long-term (2-year) follow-up, a rare occurrence in behavioral research. In essence these three studies together fail to support the differential effectiveness of complex systematic desensitization treatment packages for either severe or more moderate fears. From L. C. Miller et al. and Hampe et al., the safest conclusion is that either behavioral or psychodynamic treatment is helpful in the short term for younger children, who in any event apparently would overcome their fears without treatment in the long (2 years) term.

The articles by Leitenberg and Callahan (1973) and Kelley (1976) are clinical analogs focused on nursery school and kindergarten children selected for their fear of the dark. Both studies were well designed and used control groups, duration of dark tolerance as pre- and posttreatment behavioral measures, and visual "fear thermometers" for subjective reports by or feedback to the subjects. Neither study employed a follow-up, and no attempt was made to assess the clinical severity of the fears, leaving the possibility that the subjects were only mildly to moderately fearful.

Leitenberg and Callahan used a "reinforcement practice" method in which children were instructed to remain in a dark room for longer periods of time over five trials per session and two practice sessions per week for a maximum of 4 weeks or until the child reached a criterion of two consecutive 5-minute dark-tolerance trials. For each suc-

cessful trial the children were reinforced with praise and prizes. The authors reported a significant treatment and control group difference on posttraining dark-tolerance tests. However, there were only seven children in each group, and the mean posttraining dark tolerance for the experimental group was only about 3 minutes. Although statistically significant, does a 3-minute dark tolerance have any personal or clinical significance?

Kelley (1976) assigned 40 children who feared the dark to three desensitization treatment groups, one "play placebo" control, and one no-treatment control group and reported no differences between treatment and controls or among the three treatment conditions. The most important finding of this study is the significant effect of a simple demand manipulation, that is, verbal instructions to remain longer in the dark room dramatically increased dark tolerance and was "a far more powerful influence on both behavioral and self report change scores than three sessions of therapy" (Kelley, 1976, p. 80). In light of Kelley's findings and their implications, we suggest that Leitenberg and Callahan's "significant" reinforced practice effects may have been due simply to their verbal instructions to the experimental group only to remain in the dark room for a longer time on each subsequent trial. Kelley's finding is particularly important because it seems reasonable to hypothesize that demand characteristics are especially salient to children who, because of their immature status, are often in the position of having to comply with the outright and sometimes subtle demands of adults. Certainly these issues need to be explored. The critical research has not yet been done, but it may be that dark-tolerance performance of dark-fearful children can be significantly improved by instruction alone. But would this manipulation, perhaps over repeated trials, change other aspects of fear such as physiological and cognitive responses, and would those changes hold up over time? A potentially important area in need of research is the effects on children's fear behavior of adults' instructions.

In two studies, Mann and Rosenthal (1969) and Mann (1972) compared direct and vicarious desensitization procedures in the reduction of adolescents' text anxiety. Like subjects in Obler and Terwilliger (1970) and some of those in L. C. Miller et al. (1972a), these adolescents, referred by a school counselor for high test anxiety, appear to have been severely fearful. Their major finding was systematic desensitization, whether applied individually or in groups, whether actually or vicariously experienced, was significantly more effective in reducing test anxiety than the waiting-list control condition. The control subjects, when treated later, also improved.

Most interesting about Mann's studies is that the most effective procedure was a vicarious desensitization condition, which actually sounds very much like a symbolic modeling procedure in which subjects observe a videotaped peer model undergoing systematic desensitization and gaining control over the fear. The subjects in the vicarious condition showed greater improvement than those who actually experienced direct systematic desensitization! In Mann's work, then, although called a *desensitization* procedure, the symbolic modeling approach was the most effective with apparently highly fearful subjects. In effect, these two studies of desensitization support the effectiveness of symbolic modeling.

In three modeling studies of shy or withdrawn children, investigators tested whether the addition of direct shaping procedures added any power to the basic symbolic modeling procedure. Evers and Schwarz (1973) found that it did not. O'Connor (1972) reported that although shaping (with praise and attention) was as effective as symbolic modeling, the new interactive behavior of the shaping group soon decayed, whereas modeling group subjects maintained their gains. Somewhat similar findings were reported by Walker and Hops (1973). In their study, the addition of contingency management (tokens for increased classroom interaction) following the symbolic modeling film reliably increased social behavior beyond that produced by modeling alone. However, as Gelfand (1978) pointed out, when tokens were withdrawn, the new behavior degenerated. Apparently, "natural" reinforcers in the classroom did not maintain it. Perhaps a more gradual thinning of reinforcements or the use of more social

reinforcements would have produced a more stable level of interaction. The point here is that the success of contingency management techniques in reducing withdrawn or shy social behavior is still uncertain.

In summary, seven studies have examined complex desensitization or contingency management treatment packages in the reduction of children's fears, and several other studies have included contingency management procedures in addition to basic symbolic modeling approaches. Overall their findings are equivocal. One study (L. C. Miller et al., 1972a) found no overall beneficial effects of systematic desensitization compared with other treatments and with nontreated control conditions. Obler and Terwilliger's (1970) study has a number of methodological problems that cast doubt on their conclusions. Kelley (1976) reported no effects of desensitization treatment compared with two control treatments. Leitenberg and Callahan's (1973) results are based on a very small number of subjects and, as discussed earlier, are open to the interpretation that treatment differences may have been due simply to direct instructions. Mann and Rosenthal (1969) and Mann (1972) reported good results with a variety of systematic desensitization treatments, but, on close inspection, their most successful treatment appears to have been symbolic modeling rather than systematic desensitization. Finally, three of the modeling studies (Evers & Schwarz, 1973; O'Connor, 1972; Walker & Hops, 1973) found that the addition of contingency management did not usually improve performance beyond that achieved by symbolic modeling alone, and if it did, the new performance was not maintained.

Unlike in the modeling literature, there exists no convincing evidence that approaches developed on respondent-based systematic desensitization or operant contingency management paradigms are effective methodologies for reducing children's fears.

Cognitive Approaches

Only one published study (Kanfer et al., 1975) used a cognitive self-control approach in child fear reduction. Children 5 and 6 years old rehearsed one of three verbal-mediation responses: (a) sentences emphasizing the child's control or competence (e.g., "I am a brave girl (boy). I can take care of myself."); (b) sentences aimed at reducing the fear-stimulus value of the dark (e.g., "The dark is a fun place to be"); or (c) neutral sentences (e.g., "Mary had a little lamb"). In dark-tolerance posttests the "competence" group significantly outperformed the others. These children were not representative of a normal population, and they appeared to be only mildly or moderately fearful; we thus cannot generalize their results to clinical populations. The laboratory dark-tolerance test may have had no relationship to real fears in the children's natural environments, and no behavioral maintenance data were provided. However, despite some shortcomings it was a well-executed study with strong results. The operation of verbal self-control processes was suggested by the authors, although they noted that their study did not specify the details of verbal control mechanisms.

Other hints of the possible effectiveness of verbal mediation are found in case studies (Ayer, 1973; Lazarus & Abramovitz, 1962), in which children rehearsed highly competent, although very imaginative, strategies for coping with fear stimuli. The modeling research by Jakibchuk and Smeriglio (1976) suggests the necessity of a first-person self-speech narrative as an accompaniment to a symbolic modeling film. One possible interpretation of these various hints is that the children's own verbal self-instructions—whether originally received from a film, from specific sentences given to the subjects to rehearse, or from fanciful verbal and pictorial images—on how to successfully deal with the fear stimulus are a central component of the successful interventions. Although there are yet few appropriate data on the success of verbal self-controlling manipulations in child fear reduction, certainly enough data and hints are available to alert one that more research in this direction is indicated.

Summary of Controlled Studies

In summary, experimental fear-reduction research with children has been meager until the past 10 years, during which at least

28 studies have been published. Because of limited sampling procedures, their results cannot be generalized to either normal or clinical populations of children. As for the latter group, we have virtually no data and know surprisingly little about reducing severe phobic levels of children's fears. It has yet to be demonstrated whether the experimental fear-reduction findings are useful in clinical practice.

Only a limited number of different fears have been studied: fear of animals, fear of bodily injury (e.g., fear of dental treatment), fear of darkness, test anxiety, and social isolation. The levels of fear intensity have also been limited, with only 4 of the 28 studies including children with more than mild-to-moderate fears. Only 2 of the studies considered pretreatment fear duration, as a measure of seriousness. With the exception of dental/medical fears we have virtually no data on fear reduction outside of the laboratory setting and thus do not know if statistically significant changes brought about in the laboratory are related to fear behavior in the child's natural environment or are of sufficient magnitude to be psychologically or clinically significant. Although the experimental reduction of fear behavior has been repeatedly demonstrated, neither generalization nor long-term maintenance of the new behavior has been demonstrated.

Three major fear-reduction strategies have been employed: (a) Most studies (20 of the 28) have used modeling approaches; (b) 7 have used a variety of systematic desensitization or contingency management treatment packages; and (c) 1 study used a cognitive verbal-mediation strategy. Of the three strategies, the most reliably successful has been modeling (in terms of replications and fairly well-controlled methodology). A number of important research issues in modeling remain to be investigated, including model–subject similarity, coping versus mastery strategies, modeling effectiveness with severe fears, adequate demonstration of both generalization and long-term maintenance of the new behavior in the child's natural environment, and use of modeling in the prevention of fears.

A particularly important issue, not yet systematically investigated, is the operation of cognitive self-controlling mechanisms in successful modeling. Only one study used a verbal-mediation, self-control strategy. Its success is encouraging and should generate a good deal of research.

Unlike in the modeling literature, there is no convincing evidence that systematic desensitization or contingency management strategies are effective in reducing children's fears and in maintaining that reduction. Hatzenbuehler and Schroeder's (1978) recent review came to the same conclusion regarding desensitization procedures.

One of the more interesting, if still remote, practical implications of the recent research is the potential use of symbolic modeling in large-scale community mental health fear-prevention programs. A number of common, vexing childhood fears (e.g., fear of starting school and fear of dental/medical procedures) might be prevented on a large scale.

Overall, fear reduction and prevention in children, particularly using modeling and cognitive strategies, is a rich area for research and development. Many needed research directions in this field are clearly indicated.

Conclusions

Reaching back almost 60 years, the study of children's fears is old in the history of modern psychology. However, what is most striking is not the field's age but its apparent lack of progress. To paraphrase Berecz's (1968) conclusion, the literature still gives us only hints about the nature of children's fears. After what appears to have been a good start in the 1920s, the field experienced a long hiatus, which appears to be ending: The research of the past 10 years has returned to the behavior modification focus of the 1920s and early 1930s.

This review suggests that children's fears, like other human reactions, proceed through complex processes. The many variables involved and their interactions result in a complexity that provides many points at which to focus research. Briefly, fear stimuli may be internal, external, or both and may vary in content, number, intensity, and duration. The child's responses involve combinations of physiological, cognitive, and overt behavioral events, all of which may vary in latency,

intensity, and duration and with changes in stimulus conditions. The child's responses, overt or covert, may act on any of the stimulus and response variables, modify them, and thus occasion change in any parts of the process. All of these processes are immersed in social settings that contribute further sources of variation.

Despite this rich complexity, the research over nearly 60 years has focused almost exclusively on the endpoints of the paradigm; that is, normative research has tried to identify fear stimuli, whereas the more recent fear-reduction research has tried to modify the overt avoidance behavior. Virtually all of the variables between, observable or inferred, have been left unexplored, leaving one with little understanding of fear processes in children.

Future research will do well to follow the literature's hints and look more closely at the unexplored areas identified in this review, such as the possible adaptive value of children's fears, fear-prevention strategies, cognitive self-control variables, and developmental factors. In short, we must recognize and test far more complex paradigms of the fear process. And finally, we clearly must correct our continued lack of information regarding severe levels of children's fear.

Reference Note

1. Garcia, K. *Fears and phobias in childhood.* Unpublished manuscript, State University of New York at Buffalo, 1977.

References

Abe, K. Phobias and nervous symptoms in childhood and maturity: Persistence and associations. *British Journal of Psychiatry*, 1972, *120*, 275–283.

Adelson, R., & Goldfried, M. R. Modeling and the fearful child patient. *Journal of Dentistry for Children*, 1970, *37*, 476–489.

Adelson, R., Liebert, R. M., Poulos, R. W., & Herskovitz, A. A modeling film to reduce children's fears of dental treatment. *Journal of Dental Research*, 1972, *51*, 1708. (Abstract)

Angelino, H., Dollins, J., & Mech, E. V. Trends in the "fears and worries" of schoolchildren as related to socio-economic status and age. *Journal of Genetic Psychology*, 1956, *89*, 263–276.

Angelino, H., & Shedd, C. Shifts in the content of fears and worries relative to chronological age. *Proceedings of the Oklahoma Academy of Science*, 1953, *34*, 180–186.

Ayer, W. A. Use of visual imagery in needle-phobic children. *Journal of Dentistry for Children*, 1973, *40*, 125–127.

Ayllon, T., Smith, D., & Rogers, M. Behavioral management of school phobia. *Journal of Behavior Therapy and Experimental Psychiatry*, 1970, *1*, 125–138.

Bamber, J. H. The fears of adolescents. *Journal of Genetic Psychology*, 1974, *125*, 127–140.

Bandura, A. *Principles of behavior modification.* New York: Holt, Rinehart & Winston, 1969.

Bandura, A., & Barab, P. G. Processes governing disinhibitory effects through symbolic modeling. *Journal of Abnormal Psychology*, 1973, *82*, 1–9.

Bandura, A., Grusec, J. E., & Menlove, F. L. Vicarious extinction of avoidance behavior. *Journal of Personality and Social Psychology*, 1967, *5*, 16–23.

Bandura, A., Jeffery, R. W., & Wright, C. L. Efficacy of participant modeling as a function of response induction aids. *Journal of Abnormal Psychology*, 1974, *83*, 56–64.

Bandura, A., & Menlove, F. L. Factors determining vicarious extinction of avoidance behavior through symbolic modeling. *Journal of Personality and Social Psychology*, 1968, *8*, 99–108.

Bandura, A., Ross, D., & Ross, S. Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology*, 1973, *63*, 575–582.

Baron, R. A. Aggression as a function of magnitude of victim's pain cues, level of prior anger arousal, and aggressor-victim similarity. *Journal of Personality and Social Psychology*, 1971, *18*, 48–54.

Bauer, D. H. An exploratory study of developmental changes in children's fears. *Journal of Child Psychology and Psychiatry*, 1976, *17*, 69–74.

Begelman, D. A., & Hersen, M. Critique of Obler and Terwilliger's "Systematic desensitization with neurologically impaired children with phobic disorders." *Journal of Consulting and Clinical Psychology*, 1971, *37*, 10–13.

Bentler, P. M. An infant's phobia treated with reciprocal inhibition therapy. *Journal of Child Psychology and Psychiatry*, 1962, *3*, 185–189.

Berecz, J. M. Phobias of childhood: Etiology and treatment. *Psychological Bulletin*, 1968, *70*, 694–720.

Blanchard, E. B. Relative contributions of modeling, informational influences, and physical contact in extinction of phobic behavior. *Journal of Abnormal Psychology*, 1970, *76*, 55–61.

Bowlby, J. *Attachment and loss* (Vol. 2). New York: Basic Books, 1973.

Brown, R. E., Copeland, R. E., & Hall, R. V. School phobia: Effects of behavior modification treatment applied by an elementary school principal. *Child Study Journal*, 1974, *4*, 125–133.

Chapel, J. C. Treatment of a case of school phobia by reciprocal inhibition. *Canadian Psychiatric Association Journal*, 1967, *12*, 25–28.

Chazan, M. School phobia. *British Journal of Educational Psychology*, 1962, *32*, 209–217.

Cooper, J. A. Application of the consultant role to parent-teacher management of school avoidance behavior. *Psychology in the Schools*, 1973, *10*, 259–262.

Croake, J. W. Fears of children. *Human Development*, 1969, *12*, 239–247.
Croake, J. W., & Knox, F. H. The changing nature of children's fears. *Child Study Journal*, 1973, *3*, 91–105.
Croghan, L., & Musante, G. J. The elimination of a boy's high building phobia by *in vivo* desensitization and game playing. *Journal of Behavior Therapy and Experimental Psychiatry*, 1975, *6*, 87–88.
Cummings, J. D. The incidence of emotional symptoms in school children. *British Journal of Educational Psychology*, 1944, *14*, 151–161.
Cummings, J. D. A follow-up study of emotional symptoms in school children. *British Journal of Educational Psychology*, 1946, *16*, 163–177.
Edlund, C. V. A reinforcement approach to the elimination of a child's school phobia. *Mental Hygiene*, 1971, *55*, 433–436.
Evers, W. L., & Schwarz, J. C. Modifying social withdrawal in preschoolers: The effects of filmed modeling and teacher praise. *Journal of Abnormal Child Psychiatry*, 1973, *1*, 248–256.
Freeman, B. J., Roy, R. R., & Hemmick, S. Extinction of a phobia of physical examination in a seven year old mentally retarded boy—A case study. *Behaviour Research and Therapy*, 1976, *14*, 63–64.
Garvey, W. P., & Hegrenes, J. R. Desensitization techniques in the treatment of school phobia. *American Journal of Orthopsychiatry*, 1966, *36*, 147–152.
Geer, J. The development of a scale to measure fear. *Behaviour Research and Therapy*, 1965, *3*, 45–53.
Gelfand, D. M. Behavioral treatment of avoidance, social withdrawal and negative emotional states. In B. B. Wolman, J. Egan, & A. O. Ross (Eds.), *Handbook of mental disorders in childhood and adolescence*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.
Graziano, A. M. Reduction of children's fears. In A. M. Graziano (Ed.), *Behavior therapy with children* (Vol. 2). Chicago: Aldine, 1975.
Graziano, A. M., & DeGiovanni, I. S. The clinical significance of childhood phobias: A note on the proportion of child-clinical referrals for the treatment of children's fears. *Behaviour Research and Therapy*, 1979, *17*, 161–162.
Grossberg, J. M., & Wilson, H. K. A correlational comparison of the Wolpe-Lang Fear Survey Schedule and the Taylor Manifest Anxiety Scale. *Behaviour Research and Therapy*, 1965, *3*, 125–128.
Hagman, E. A study of fears of children of preschool age. *Journal of Experimental Education*, 1932, *1*, 110–130.
Hampe, E., Noble, H., Miller, L. C., & Barrett, C. L. Phobic children one and two years posttreatment. *Journal of Abnormal Psychology*, 1973, *82*, 446–453.
Hatzenbuehler, L. C., & Schroeder, H. E. Desensitization procedures in the treatment of childhood disorders. *Psychological Bulletin*, 1978, *85*, 831–844.
Herrnstein, R. J. Method and theory in the study of avoidance. *Psychological Review*, 1969, *76*, 49–69.
Hersen, M. Behavior modification approach to a school phobic case. *Journal of Clinical Psychology*, 1970, *26*, 128–132.
Hersen, M. The behavioral treatment of school phobia. *Journal of Nervous and Mental Disease*, 1971, *153*, 99–107.
Hicks, D. Imitation and retention of film-mediated aggressive peer and adult models. *Journal of Personality and Social Psychology*, 1965, *2*, 97–100.
Hill, J., Liebert, R., & Mott, D. Vicarious extinction of avoidance behavior through film: An initial test. *Psychological Reports*, 1968, *12*, 192.
Holmes, F B. An experimental investigation of a method of overcoming children's fears. *Child Development*, 1936, *7*, 6–30.
Jakibchuk, Z., & Smeriglio, V. L. The influence of symbolic modeling on the social behavior of preschool children with low levels of social responsiveness. *Child Development*, 1976, *47*, 838–841.
Jersild, A. T. *Child psychology* (6th ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1968.
Jersild, A. T., & Holmes, F. B. Children's fears. *Child Development Monograph*, No. 20, 1935. (a)
Jersild, A. T., & Holmes, F. B. Methods of overcoming children's fears. *Journal of Psychology*, 1935, *1*, 75–104. (b)
Jersild, A. T., Markey, F. U., & Jersild, C. L. Children's fears, dreams, wishes, daydreams, likes, dislikes, pleasant and unpleasant memories. *Child Development Monograph*, No. 12, 1933.
Johnson, R., & Machen, J. B. Behavior modification techniques and maternal anxiety. *Journal of Dentistry for Children*, 1973, *40*, 272–276.
Jones, M. C. The elimination of children's fears. *Journal of Experimental Psychology*, 1924, *7*, 382–390. (a)
Jones, M. C. A laboratory study of fear: The case of Peter. *Pedagogical Seminar*, 1924, *31*, 308–315. (b)
Kanfer, F. H., Karoly, P., & Newman, A. Reduction of children's fear of the dark by competence-related and situational threat-related verbal cues. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 251–258.
Katz, R. C. Single session recovery from a hemodialysis phobia: A case study. *Journal of Behavior Therapy and Experimental Psychiatry*, 1974, *5*, 205–206.
Keller, M. F., & Carlson, P. M. The use of symbolic modeling to promote social skills in preschool children with low levels of social responsiveness. *Child Development*, 1974, *45*, 912–919.
Kelley, C. K. Play desensitization of fear of darkness in preschool children. *Behaviour Research and Therapy*, 1976, *14*, 79–81.
Kelly, E. W. School phobia: A review of theory and treatment. *Psychology in the Schools*, 1973, *10*, 33–42.
Kennedy, W. A. School phobia: Rapid treatment of fifty cases. *Journal of Abnormal Psychology*, 1965, *70*, 285–289.
Kissel, S. Systematic desensitization therapy with children: A case study and some suggested modifications. *Professional Psychology*, 1972, *3*, 164–168.
Kornhaber, R. C., & Schroeder, H. E. Importance of model similarity on the extinction of avoidance behavior in children. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 601–607.
Lang, P. J., & Lazovik, A. D. The experimental

desensitization of an animal phobia. *Journal of Abnormal and Social Psychology*, 1963, *66*, 519–525.

Lapouse, R., & Monk, M. A. Fears and worries in a representative sample of children. *American Journal of Orthopsychiatry*, 1959, *29*, 803–818.

Lazarus, A. A. The elimination of children's phobias by deconditioning. In H. J. Eysenck (Ed.), *Behavior therapy and the neuroses*. New York: Pergamon Press, 1960.

Lazarus, A. A., & Abramovitz, A. The use of "emotive imagery" in the treatment of children's phobias. *Journal of Mental Science*, 1962, *108*, 191–195.

Leitenberg, H., & Callahan, E. J. Reinforced practice and reduction of different kinds of fears in adults and children. *Behaviour Research and Therapy*, 1973, *11*, 19–30.

Lewis, S. A comparison of behavior therapy techniques in the reduction of fearful avoidance behavior. *Behavior Therapy*, 1974, *5*, 648–655.

Maccoby, F. F., & Wilson, W. C. Identification and observational learning from films. *Journal of Abnormal and Social Psychology*, 1957, *55*, 76–87.

MacFarlane, J. W., Allen, L., & Honzik, M. P. *A developmental study of the behavior problems of normal children between twenty-one months and fourteen years*. Berkeley: University of California Press, 1954.

Machen, J. B., & Johnson, R. Desensitization, model learning, and the dental behavior of children. *Journal of Dental Research*, 1974, *53*, 83–87.

Mann, J. Vicarious desensitization of test anxiety through observation of videotaped treatment. *Journal of Counseling Psychology*, 1972, *19*, 1–7.

Mann, J., & Rosenthal, T. L. Vicarious and direct counterconditioning of test anxiety through individual and group desensitization. *Behaviour Research and Therapy*, 1969, *7*, 359–367.

Manosevitz, M., & Lanyon, R. I. Fear Survey Schedule: A normative study. *Psychological Reports*, 1965, *17*, 699–703.

Marks, I. M. *Fears and phobias*. New York: Academic Press, 1969.

Marks, I. M., & Gelder, M. G. Different ages of onset in varieties of phobia. *American Journal of Psychiatry*, 1966, *123*, 218–221.

Mathews, A. Fear-reduction research and clinical phobias. *Psychological Bulletin*, 1978, *85*, 390–404.

Maurer, A. What children fear. *Journal of Genetic Psychology*, 1965, *106*, 265–277.

Melamed, B. G., Hawes, R. R., Heiby, E., & Glick, J. The use of filmed modeling to reduce uncooperative behavior of children during dental treatment. *Journal of Dental Research*, 1975, *54*, 797–801.

Melamed, B. G., & Siegel, L. J. Reduction of anxiety in children facing hospitalization and surgery by use of filmed modeling. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 511–521.

Melamed, B. G., Weinstein, D., Hawes, R., & Katin-Borland, M. Reduction of fear-related dental management problems using filmed modeling. *Journal of the American Dental Association*, 1975, *90*, 822–826.

Meichenbaum, D. Examination of model characteristics in reducing avoidance behavior. *Journal of Personality and Social Psychology*, 1971, *17*, 298–307.

Meichenbaum, D., & Turk, D. The cognitive-behavioral management of anxiety, anger, and pain. In P. O. Davidson (Ed.), *The behavioral management of anxiety, depression and pain*. New York: Brunner/Mazel, 1976.

Miklich, D. R. Operant conditioning procedures with systematic desensitization in a hyperkinetic asthmatic boy. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, *4*, 177–182.

Miller, L. C. Louisville Behavior Check List for males 6–12 years of age. *Psychological Reports*, 1967, *21*, 885–896.

Miller, L. C., Barrett, C. L., & Hampe, E. Phobias of childhood in a prescientific era. In A. Davids (Ed.), *Child personality and psychopathology: Current topics* (Vol. 1). New York: Wiley, 1974.

Miller, L. C., Barrett, C. L., Hampe, E., & Noble, H. Revised anxiety scales for the Louisville Behavior Check List. *Psychological Reports*, 1971, *29*, 503–511.

Miller, L. C., Barrett, C. L., Hampe, E., & Noble, H. Comparison of reciprocal inhibition, psychotherapy, and waiting list control for phobic children. *Journal of Abnormal Psychology*, 1972, *79*, 269–279. (a)

Miller, L. C., Barrett, C. L., Hampe, E., & Noble, H. Factor structure of childhood fears. *Journal of Consulting and Clinical Psychology*, 1972, *39*, 264–268. (b)

Miller, P. M. The use of visual imagery and muscle relaxation in the counter-conditioning of a phobic child: A case study. *Journal of Nervous and Mental Disease*, 1972, *154*, 457–460.

Morgan, G. A. V. Children who refuse to go to school. *Medical Officer*, 1959, *102*, 221–224.

Mowrer, O. H. A stimulus-response analysis of anxiety and its role as a reinforcing agent. *Psychological Review*, 1939, *46*, 553–565.

Murphy, C. M., & Bootzin, R. R. Active and passive participation in the contact desensitization of snake fear in children. *Behavior Therapy*, 1973, *4*, 203–211.

Nalven, F. B. Manifest fears and worries of ghetto versus middle class suburban children. *Psychological Reports*, 1970, *27*, 285–286.

Newstatter, W. L. The effect of poor social conditions in the production of neurosis. *Lancet*, 1938, *234*, 1436–1441.

Ney, P. G. Combined psychotherapy and deconditioning of a child's phobia. *Canadian Psychiatric Association Journal*, 1968, *13*, 293–294.

Obler, M., & Terwilliger, R. F. Pilot study on the effectiveness of systematic desensitization with neurologically impaired children with phobic disorders. *Journal of Consulting and Clinical Psychology*, 1970, *34*, 314–318.

O'Connor, R. D. Modification of social withdrawal through symbolic modeling. *Journal of Applied Behavior Analysis*, 1969, *2*, 15–22.

O'Connor, R. D. Relative efficacy of modeling, shaping, and the combined procedures for modification of social withdrawal. *Journal of Abnormal Psychology*, 1972, *79*, 327–334.

Ollendick, T. H., & Gruen, G. E. Treatment of a bodily injury phobia with implosive therapy. *Journal of Consulting and Clinical Psychology*, 1972, *38*, 389–393.

Olsen, I., & Coleman, H. S. Treatment of school phobia as a case of separation anxiety. *Psychology in the Schools*, 1967, *4*, 151–154.

Patterson, G. R. A learning theory approach to the treatment of the school phobic child. In L. P. Ullman & L. Krasner (Eds.), *Case studies in behavior modification*. New York: Holt, Rinehart & Winston, 1965.

Piaget, J. *The child's conception of the world*. New York: Harcourt, Brace, 1929.

Piaget, J. *The child's conception of physical causality*. New York: Humanities Press, 1951.

Poser, E. G., & King, M. C. Strategies for the prevention of maladaptive fear response. *Canadian Journal of Behavioral Science*, 1975, *7*, 279–294.

Poznanski, E. O. Children with excessive fears. *American Journal of Orthopsychiatry*, 1973, *43*, 428–438.

Pratt, K. C. The study of the "fears" of rural children. *Journal of Genetic Psychology*, 1945, *67*, 179–194.

Rachman, S. Clinical applications of observational learning, imitation and modeling. *Behaviour Research and Therapy*, 1972, *3*, 379–397.

Rachman, S. The passing of the two-stage theory of fear and avoidance: Fresh possibilities. *Behaviour Research and Therapy*, 1976, *14*, 125–134.

Rachman, S. The conditioning theory of fear-acquisition: A critical examination. *Behaviour Research and Therapy*, 1977, *15*, 375–387.

Rachman, S. *Fear and courage*. San Francisco- Freeman, 1978.

Rachman, S., & Costello, C. G. The aetiology and treatment of children's phobias—A review. *American Journal of Psychiatry*, 1961, *118*, 97–105.

Rachman, S., & Hodgson, R. I. Synchrony and desynchrony in fear and avoidance. *Behaviour Research and Therapy*, 1974, *12*, 311–318.

Rhines, W. B. Behavior therapy before institutionalization. *Psychotherapy: Theory, Research & Practice*, 1973, *10*, 281–283.

Rimm, D. C., & Masters, J. C. *Behavior therapy: Techniques and empirical findings*. New York: Academic Press, 1974.

Ritter, B. The group desensitization of children's snake phobias using vicarious and contact desensitization procedures. *Behaviour Research and Therapy*, 1968, *6*, 1–6.

Rodriguez, A., Rodriguez, M., & Eisenberg, L. The outcome of school phobia. *American Journal of Psychiatry*, 1959, *116*, 540–544.

Ross, D. M. Effect on learning of psychological attachment to a film model. *American Journal of Mental Deficiency*, 1970, *74*, 701–707.

Russell, G. W. Human fears: A factor analytic study of three age levels. *Genetic Psychology Monographs*, 1967, *76*, 141–162.

Rutter, M., Tizard, J., & Whitmore, K. *Education, health and behaviour*. New York: Wiley, 1970.

Scherer, M. W., & Nakamura, C. Y. A Fear Survey Schedule for Children (FSS-FC): A factor analytic comparison with manifest anxiety (CMAS). *Behaviour Research and Therapy*, 1968, *6*, 173–182.

Shepherd, M., Oppenheim, B., & Mitchell, S. *Childhood behaviour and mental health*. London: University of London Press, 1972.

Smith, R. E., & Sharpe, T. O. Treatment of a school phobia with implosive therapy. *Journal of Consulting and Clinical Psychology*, 1970, *35*, 239–243.

Solomon, R. L., & Turner, L. H. Discriminative classical conditioning in dogs paralyzed by curare can later control discriminative avoidance responses in the normal state. *Psychological Review*, 1962, *69*, 202–219.

Spiegler, M., & Liebert, R. Some correlates of self-reported fear. *Psychological Reports*, 1970, *26*, 691–695.

Spiegler, M., Liebert, R., McMains, M., & Fernandez, L. Experimental development of a modeling treatment to extinguish persistent avoidance behavior. In R. D. Rubin & C. Franks (Eds.), *Advances in behavior therapy, 1968*. New York: Academic Press, 1969.

Tahmisian, J. A., & McReynolds, W. Use of parents as behavioral engineers in the treatment of a school-phobic girl. *Journal of Counseling Psychology*, 1971, *18*, 225–228.

Tasto, D. Systematic desensitization, muscle relaxation and visual imagery in the counter-conditioning of a four year old phobic child. *Behaviour Research and Therapy*, 1969, *7*, 409–411.

Ullmann, L. P., & Krasner, L. *A psychological approach to abnormal behavior* (2nd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Vaal, J. J. Applying contingency contracting to a school phobic: A case study. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, *4*, 371–373.

Vernon, D. T. A., & Bailey, W. C. The use of motion pictures in the psychological preparation of children for induction of anesthesia. *Anesthesiology*, 1974, *40*, 68–72.

Walker, H. M., & Hops, H. Group and individual reinforcement contingencies in the modification of social withdrawal. In L. A. Hamerlynck, L. Handy, & E. Mash (Eds.), *Behavior change: Methodology, concepts and practice*. Champaign, Ill.: Research Press, 1973.

Watson, J. B., & Rayner, R. Conditioned emotional reactions. *Journal of Experimental Psychology*, 1920, *3*, 1–14.

Weber, H. An approach to the problem of fear in children. *Journal of Mental Science*, 1936, *82*, 136–147.

Weinberger, G., Leventhal, T., & Beckman, G. The management of a chronic school phobic through the use of consultation with school personnel. *Psychology in the Schools*, 1973, *10*, 83–88.

Wheeler, L., & Levine, L. Observer-model similarity in the contagion of aggression. *Sociometry*, 1967, *30*, 41–49.

White, W. C., Jr., & Davis, M. T. Vicarious extinction of phobic behavior in early childhood. *Journal of Abnormal Child Psychology*, 1974, *2*, 25–32.

Wilson, G. D. Social desirability and sex differences

in expressed fear. *Behaviour Research and Therapy*, 1967, *5*, 136–137.

Wish, P. A., Hasazi, J. E., & Jurgela, A. R. Automated direct deconditioning of a childhood phobia. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, *4*, 279–283.

Wolpe, J. *The practice of behavior therapy*. Elmsford, N.Y.: Pergamon Press, 1974.

Wolpe, J., Brady, J. P., Serber, M., Agras, W. S., & Liberman, R. P. The current status of systematic desensitization. *American Journal of Psychiatry*, 1973, *130*, 961–965.

Yates, A. J. *Behavior therapy*. New York: Wiley, 1970.

Received March 27, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 831–851

Cognitive Development in Retarded and Nonretarded Persons: Piagetian Tests of the Similar Sequence Hypothesis

John R. Weisz
University of North Carolina at Chapel Hill

Edward Zigler
Yale University

From the debate over developmental "universals" in Piagetian theory and the controversy between developmental and difference theories of mental retardation, an important hypothesis emerges—one that is testable via cognitive–developmental comparisons between retarded and nonretarded persons. This *similar sequence hypothesis* holds that retarded and nonretarded persons traverse the same stages of cognitive development in the same order, differing only in the rate at which they progress and in the ultimate developmental ceiling they attain. Current evidence relevant to this hypothesis is drawn from 3 longitudinal and 28 cross-sectional studies of developmental phenomena described by Piaget. The great preponderance of this evidence supports the hypothesis with respect to every subject group, with the possible exception of individuals suffering from pronounced electroencephalogram abnormalities. The quality of current evidence is critically evaluated, and procedures by which more precise tests of the hypothesis might be fashioned are proposed. Overall, the review illustrates that developmental research with atypical populations can be a potent tool in testing general developmental theory. Conversely, it illustrates the power of general developmental theory to enrich our understanding of atypical development.

In recent years two important theoretical issues have stimulated interest in Piagetian research with retarded and nonretarded populations. One is the question of whether developmental "universals" exist. Many psychologists regard the sequence of developmental stages described by Piaget (e.g., 1970) and elaborated by other cognitive–developmental theorists (e.g., Kohlberg, 1969) as one of psychology's few current candidates for universality (see Weisz, 1978). Piaget (1956, 1966) took a psychological universalist position, with qualifications, and Kohlberg (1969, 1971) argued strongly for the invariance of what he regarded as a cognitive–developmental sequence rooted in an inherent logic and in universal characteristics of both the nervous system and the environment.

Of course it is impossible to know that any given developmental phenomenon occurs everywhere without exception, since one can never test all possible exceptions (Popper, 1959). However, if one is not to let the claims of cognitive–developmental theorists go unchallenged, it is important to evaluate the extent to which *transcontextual validity* (see Weisz, 1978) has been demonstrated for the Piagetian account of development.

One approach to assessing such validity across experimental contexts is to examine developmental sequences across various cultures (Buck-Morss, 1975; Simpson, 1974). Another approach, of particular interest because of the cognitive emphasis of Piagetian theory, is to compare groups of children who differ markedly in measured intelligence, that is, groups of mentally retarded and nonretarded children. If children at very different IQ levels were to show identical Piagetian developmental sequences, then the transcon-

This review was supported by a New York State Human Ecology research grant to the first author and by National Institute of Child Health and Human Development Grant 5 POI HD 03008 to the second author. The authors are grateful to Thomas Achenbach and Sally Styfco for their thoughtful comments on a draft of the article.

Requests for reprints should be sent to John R. Weisz, Department of Psychology, University of North Carolina, Chapel Hill, North Carolina 27514.

textual validity of the Piagetian account of development would be substantially supported. If retarded and nonretarded children were to differ in their sequence of development, then universality could hardly be claimed for the Piagetian account.

A second theoretical issue that has sparked recent interest in comparative cognitive research is reflected in the ongoing debate between proponents of *developmental* and *difference* theories of mental retardation. The developmental position, set forth by Zigler (1969), is intended to apply to retarded individuals not suffering from organic impairment. Zigler has maintained that the retarded child passes through cognitive–developmental stages in the same order as the nonretarded child, with only two differences: The retarded child passes through the stages more slowly and attains a lower upper limit relative to the nonretarded child.[1]

A number of theorists hold what Zigler labeled the *general difference* position. One aspect of this position is the view that the cognitive development of retarded persons differs from that of nonretarded persons in ways that go beyond mere differences in rate and ceiling of development. Milgram (1973), for example, maintained that the cognitive levels, or stages, of retarded children differ from those of the nonretarded in that the former are more likely to contain traces of developmentally earlier levels and are more likely to show regression to those earlier levels. (For further discussion of the developmental position, the specific difference positions, and the rationale underlying them, see Weisz & Zigler, in press.)

This theoretical conflict has generated a new emphasis on comparative research into the processes (rather than the products) of learning and reasoning (Weisz, 1977; Weisz & Achenbach, 1975) and into processes of reasoning, as described in Piagetian theory (see Wilton & Boersma, 1974).

The growing interest in the pursuit of developmental universals, and the growing intensity of the developmental versus difference debate, have thus combined to lend theoretical force to research comparing the cognitive development of retarded and nonretarded persons along Piagetian lines. This body of research has grown rapidly within the past decade; it now appears to be substantial enough to serve as a resource in our efforts to answer the principal question raised by the universality issue and the developmental versus difference debate. This question can be stated in the form of a testable hypothesis.

Similar Sequence Hypothesis

An appropriate label seems to be the *similar sequence hypothesis*. The hypothesis holds that during development retarded and nonretarded persons traverse the same stages in precisely the same order and differ only in rate of development and in the ultimate ceiling they attain. To be precise, the developmental position (Zigler, 1969, 1971) generates this hypothesis only with respect to nonretarded and cultural–familial retarded persons (thus excluding, for example, brain-damaged and genetically impaired individuals).[2] In

---

[1] An additional postulate of the developmental position is that familial retarded and nonretarded persons who are equivalent in developmental level (often operationally defined as mental age) do not differ in the formed cognitive processes they employ in learning and reasoning. This particular proposition is not germane to the present review and consequently is not discussed here. However, Piagetian evidence bearing on this proposition is being reviewed (Weisz & Zigler, in press).

[2] The reasoning underlying this qualification bears brief explanation. The developmental position holds that mental retardation can be viewed as a developmental phenomenon most appropriately among persons whose retardation does not result from specific physiological defects. Such investigators as Benton (e.g., 1962), Cruikshank (e.g., 1967), and Reitan (e.g., 1973) have devoted many years to demonstrating idiosyncratic performance characteristics that distinguish brain-injured individuals from those with intact nervous systems. Furthermore, a number of studies employing the specific kinds of problem-solving tasks most often used in research on the developmental–difference controversy have revealed effects of organicity on retarded children's performance (Balla, Styfco, & Zigler, 1971; Balla & Zigler, 1964; Elkind, Koegler, Go, & Van Doorninck, 1965; Harter, Brown, & Zigler, 1971). In harmony with such findings, proponents of the developmental position have adhered to the two-group approach (see Zigler, 1969), whereby familial retarded individuals are distinguished from those suffering from organic impairment (including genetic anomalies such as Down's syndrome). There is some disagreement among investigators over the need for the two-group approach (Ellis, 1969; Milgram,

cognitive–developmental theory, however, the claims for the universality of the developmental sequence appear to be broader. Piaget (1956) held that "the *minimum* program for establishment of stages is the recognition of a distinct chronology, in the sense of a *constant order of succession*" (p. 13). According to Kohlberg (1969), the claim that there is an invariant order of cognitive stages rests upon an assumed invariance in certain features of the environment and of the nervous system and upon "a logical analysis of orderings inherent in given concepts" (p. 355). These inherent orderings are seen as logically essential and as independent of individual differences among people. Kohlberg continued, "The invariance of sequence in the development of a concept or category is not dependent upon a prepatterned unfolding of neural patterns; it must depend upon a logical analysis of the concept itself" (p. 355). Thus, the similar sequence hypothesis as advanced by cognitive–developmental theorists seems to predict a truly universal ordering of stages—an ordering that is the same for retarded children of all etiologies (including genetic impairment, brain injury, and other neurological anomaly) as it is for all nonretarded children. There is a conservative version of the similar sequence hypothesis that applies only to familial retarded and nonretarded persons and a liberal version that applies to all persons. In the present article we present evidence bearing on both versions.

In contrast with both these versions, Milgram (1973) has argued that the retarded child's cognitive stages differ from those of the nonretarded child. In contrast with the liberal version of the hypothesis, Rogers (1977) has described a rationale for (though she has not necessarily endorsed) the hypothesis that profoundly (and thus nonfamilial) retarded children have abnormal developmental patterns.

1973). Moreover, there is a strong Piagetian rationale for applying the similar sequence hypothesis to all persons, retarded or nonretarded, organically impaired or intact (see the remainder of the paragraph for details). In what follows we describe this rationale, and we go on to review evidence in a manner that bears directly on the conservative, two-group-oriented version of the hypothesis and on the more liberal version in which the similar sequence hypothesis is applied to all retarded groups regardless of etiology.

Material Excluded From the Present Review

The present article is an attempt to synthesize studies relating to the similar sequence hypothesis. In selecting studies to be reviewed, we excluded studies of reading per se and of language development per se. Although both areas can be viewed from the perspective of Piagetian theory, neither is central to the theory; furthermore, the research in both areas is now so voluminous as to warrant separate review. We also excluded studies designed to accelerate cognitive development, since it is not the purpose of this article to determine whether retarded or nonretarded children can be trained more readily.

The studies we do include in this review vary widely in their sampling procedures, their experimental methodology, and their approaches to data analysis and reporting. Consequently, the studies differ in their level of importance vis-à-vis the hypothesis of particular interest here. For this reason, we reserve the right to vary the level of detail in which we describe the studies and give relatively greater space to those that seem to afford the clearest tests of the hypothesis.

Tests of the Similar Sequence Hypothesis

Cross-Sectional and Order-of-Difficulty Evidence

One approach to testing the similar sequence hypothesis is to assess groups of mentally retarded children at more than one developmental level with respect to their performance on various Piagetian tasks. If the direction of the difference in performance from one developmental level to another is the same for the retarded as for the nonretarded, or if the direction is consistent with the developmental sequence posited by cognitive–developmental theory, then the similar sequence hypothesis is supported. A second general approach to testing the similar sequence hypothesis is to rely less upon the developmental levels of the groups sampled than upon the relative-difficulty levels of the various tasks or behavioral items being employed. Perhaps the simplest, but also the least informative, of the variants of this approach

is to rank order the items with respect to the number of subjects who pass each one; if this rank ordering of a retarded sample matches either the rank ordering of a nonretarded sample or the developmental sequence posited by cognitive–developmental theory, then the similar sequence hypothesis is supported, albeit modestly. A more informative type of order-of-difficulty evidence is the type that employs scaling procedures, allowing one to determine, for example, how many of the children who grasp Concept A also grasp Concept B and vice versa. Such evidence, when combined with Guttman-type (e.g., Guttman, 1950) scalogram analyses, can provide a relatively strong test of the similar sequence hypothesis. The studies reviewed in this section all employed some type of cross-sectional evidence, order-of-difficulty evidence, or a combination of the two.

Development in the sensorimotor period. Early evidence bearing on the similar sequence hypothesis was provided by the research of Woodward (1959, 1961, 1962, 1963). The first of her studies (1959) focused primarily on a group of 65 institutionalized children and adolescents with a chronological age (CA) range of 7–16 years who were so profoundly retarded that they failed to attain a basal age of 2 years on the Terman-Merrill scale. Although the author maintained that this sample excluded cases involving motor or sensory disability, the cases involved a diversity of medical problems (e.g., 19 subjects were epileptic), and 38 of the children were "emotionally unstable." Woodward used three means of assessing the sensorimotor stages of this heterogeneous group. First, she observed their spontaneous mannerisms and their manipulation of toys presented individually and in a standardized order. Second, each subject was presented with three pairs of tasks, each pair tapping one of Piaget's (1953, 1955) last three sensorimotor stages (there are six stages in all). Third, Woodward presented each child with a series of object concept tasks in which a piece of candy or a toy was first used to attract the subject's attention and was then withdrawn and concealed to varying degrees.

All but the object concept tasks were analyzed in a way that sheds light on the similar sequence hypothesis. Each task was classified with respect to the Piagetian sensorimotor stage it was designed to represent; then the tasks were ranked ordered with respect to the percentage of subjects passing each. The difficulty level rankings of these 11 items were identical to the Piagetian stage level order, with one exception: A task involving coordination of vision and hearing (Sensorimotor Stage 2) proved to be slightly more difficult than a task involving manipulation of objects (Stage 3); 53 subjects passed the manipulation task, and only 49 passed the coordination task. Furthermore, when the possibly insensitive coordination task was removed from the analyses, 59 of the 65 children passed all of the items at stages below their highest stage level response. Given the extreme diversity of this sample, the high incidence of emotional instability, and the apparent tendency of many not to show responses of which they were actually capable (e.g., some delayed for a half hour before grasping an object placed before them), these data lend surprisingly strong support to the similar sequence hypothesis.

Recently, Rogers (1977) undertook an investigation similar to that of Woodward in several respects. The subjects, 40 profoundly retarded children ranging in age from 8–14 years, with IQs below 20, were given a series of Piagetian tasks. By means of these tasks, each child's performance was classified into Sensorimotor Stage 3, 4, 5, or 6 in each of four conceptual domains: object permanence (tasks involving searches for a hidden object), spatiality (tasks involving visual anticipation and rotation of objects), causality (tasks involving the use of physical prompts and tools, the removal of obstacles, and inference as to the cause of a jingling sound inside a box), and imitation (tasks involving the reproduction of both self-initiated and experimenter-initiated movements and sounds). Performance within each of the four domains was analyzed using scaling techniques, and Guttman's (1950) coefficient of reproducibility and Green's (1956) index of scalability were calculated for each scale. The object permanence and imitation tasks formed highly reproducible scales in the orders hypothesized by Piaget (1955, 1962, 1972). Causality tasks also

formed a highly reproducible scale, although the item order differed from the predicted sequence in one respect: One Stage 6 item preceded one Stage 5 item. The author attributed this irregularity to a poor choice of Stage 6 task (i.e., opening a box to obtain a bell when box opening has just been demonstrated to the subject), "since the task used might have been accomplished using imitation rather than problem-solving skills" (Rogers, 1977; pp. 841–842). Finally, the individual spatiality tasks did not all form highly reproducible scales, but when the tasks within each stage were combined (and subjects were credited with a stage level for passing one or more of the tasks from that level), the stages did form a highly reproducible scale. Rogers concluded convincingly that her findings support "the invariant sequentiality of sensorimotor stages" (p. 841).

The preoperational–concrete operational transition—the Inhelder study of conservation. One of the earliest studies bearing on the similar sequence hypothesis was carried out by Piaget's associate, Barbel Inhelder, in the early 1940s. This study, now published in English (Inhelder, 1943/1968), involved the assessment of conservation of substance, weight, and volume in 159 persons who had been labeled mentally retarded by Swiss education officials. The sample was extremely heterogeneous (see Jordan, 1976). Ages ranged from 7½–52 years, IQs ranged from 35–104, institutionalized and noninstitutionalized persons were included, and the range of etiologies and physical maladies included such diverse states as "defective environment," rickets, hearing defect, "abandoned," and schizophrenia. The procedure involved semistructured clinical interviews with each subject. Since the procedure was not perfectly standardized and little in the way of formal data analysis was presented, it is difficult to evaluate Inhelder's conclusions, including her references to "oscillations" in the reasoning of retarded subjects, discussed later in this article. However, in Piaget's (1968) description of the Inhelder study, he explained that in the entire sample,

> not one [individual] understood the conservation of weight without having the conservation of substance, nor the conservation of volume without both weight and substance, while the conservation of substance *was* found without the other two, and the conservation of weight was found without the conservation of volume. (p. 11)

Given the marked heterogeneity of the sample, such uniform support for the similar sequence hypothesis is noteworthy.

Other studies of conservation and related concepts using retarded samples only. Studies of conservation and related concepts done since the Inhelder investigation have a bearing on the similar sequence hypothesis, despite the fact that they only sampled retarded subjects. Klauss and Green (1972) assessed conservation of number and volume in 27 trainable mentally retarded subjects ranging in age from 13–19 years and in IQ from 29–57. These investigators found that volume conservation presented greater difficulty than did number conservation, a finding consistent with the pattern apparent in the nonretarded. Marchi (1971) tested conservation of mass, weight, and volume in 106 educable mentally retarded children. Difficulty level evidence suggested that contrary to Marchi's prediction, the retarded "follow a similar sequence in the acquisition of mass, weight, and volume as postulated for normals" (p. 6442).

Roodin, Sullivan, and Rybash (1976) assessed qualitative identity, quantitative identity, and equivalence conservation (see Elkind, 1967) in 60 institutionalized retarded children averaging 13 years of age and about 47 in IQ. Dyed water was poured from a standard 100 ml beaker; to test qualitative identity, subjects were asked, "Is the water in this glass (comparison) the *same* water that was in that glass (empty standard)?" To assess quantitative identity, subjects were asked, "Is there *as much* water in this glass (comparison) as there was in that glass (empty standard)?" To assess equivalence conservation, two standard beakers were filled with equal levels of water, and the contents of one were then poured into a comparison beaker; the experimenter then asked, "Is there *as much* water in this glass (standard) as there is in this glass (comparison)?" Previous reasearch (e.g., Papalia & Hooper, 1971) with nonretarded children had suggested that the developmental order for the attainment of these three concepts would be qualitative identity, quantitative identity,

and equivalence conservation. In the Roodin et al. study, analyses of the number of conservers on each task indicated a parallel order of difficulty.

In a similar study also employing 60 institutionalized retarded children (age range of approximately 10–16 years and average IQ of approximately 57), McManis (1969c) investigated identity and equivalence conservation with three types of material (Styrofoam balls, clay, and water). Like Roodin et al. (1976), McManis found evidence that the developmental sequence of his retarded subjects replicated that of nonretarded children. The notion that identity conservation must precede equivalence conservation was supported by the finding that no subject who failed to achieve identity conservation showed equivalence conservation, whereas 13%–18% of the subjects (precise percentage depending on the particular task used) displayed identity conservation without equivalence conservation.

Three studies that examined conservation of number, and number concepts generally, in mentally retarded groups yielded similar findings, despite some differences in methodology. Woodward (1961) investigated numerical concepts of 94 institutionalized individuals (50 adults with average CA of 19 years and 44 children and adolescents with average CA of 12 years) ranging in IQ from 25–73. Tests given to the subjects included assessments of their understanding of (a) one-to-one correspondence and equivalence of corresponding sets, (b) ways of equalizing unequal groups, (c) seriation, and (d) conservation of continuous quantity (water and sand). Performance was scored as indicative of one of two preoperational stages or of concrete operational thinking. When the stage level assignments were plotted as a function of the IQs (and thus roughly of the mental ages or MAs) of the adult subjects sampled, the table reflected precisely what would be expected from the application of Piaget's stage scheme to nonretarded individuals.

In the second of the three studies, Mannix (1960) administered eight of Piaget's (1952) number concept tasks to 48 "educationally subnormal" individuals ranging in MA from 5–9 years. The tasks included two tests of additive composition, one test of coordination of equivalence relations, two tests of judgment of correspondence between sets of items, and two conservation tasks (continuous and discontinuous quantities). Responses to these tests were classified into Piagetian stage levels, a scalogram was constructed, and the coefficient of reproducibility was .94. Mannix's brief report gave little information as to the precise nature of the scale types; but apparently the scalogram was consistent with Piaget's stage theory, because the author concluded that educationally subnormal children "pass through the three stages of development described by Piaget" (Mannix, 1960, p. 181).

The third of these studies on number concepts was conducted with 20 institutionalized mentally retarded persons in New Zealand (CA range of 8–17 years; IQs of 29–65). Singh and Stott (1975) presented these subjects with a series of number conservation tasks designed to classify them with respect to three Piagetian number stages: Stage 1—child fails to attend to relevant cues and fails to conserve; Stage 2—child selectively attends to only certain relevant cues, can match perceptually, but cannot conserve; Stage 3—child conserves, showing understanding of invariance of properties despite transformation in appearance. Data bearing on the similar sequence hypothesis are not reported in detail, but the authors' conclusion is quite clear: "Retarded children apparently develop sequentially in the same order as normals but at a slower rate and at a later CA" (Singh & Stott, 1975, p. 220).

One other study that used only a retarded sample deserves mention both because of its scope and because of an important issue it raises. In this study, Lister (1972) assessed six types of conservation among 115 educationally subnormal pupils in Great Britain. The subjects were aged 8–16 years, and their IQs ranged from 47–81. Both difficulty level rankings and a scaling procedure strongly suggested the following developmental sequence in the emergence of these types of conservation: number, substance, length, weight, volume, and area. Although no scalogram statistics were calculated, only 6 of the 115 subjects showed a scalogram response pattern inconsistent with the preceding order.

Lister noted that the order with respect to substance, weight, and volume was consistent with previous Piagetian research, whereas the suggested order of the remaining attributes differed from at least some previous findings with nonretarded subjects. Her own interpretation of the discrepancies was that they resulted from experiment-to-experiment variations in the specifics of the problems used to assess the various types of conservation. This is a very real possibility, and it is one reason why tests of the similar sequence hypothesis that expose retarded and nonretarded subjects to the same experimental procedures must be regarded as stronger evidence than experiments that test only retarded children and compare the findings with those of different experiments. We now turn to six studies of the former type.

Studies of conservation and related concepts employing both retarded and nonretarded subjects. Four of the experiments in which the performance of retarded and nonretarded subjects was directly compared were conducted by McManis (1969b, 1969d, 1969e, 1970). In one of these, 90 institutionalized retarded subjects (IQs of 47–73) and 90 nonretarded elementary school children (IQs of 85–115) were tested for conservation of mass, weight, and volume of clay, using Piaget and Inhelder's (1941) "sausage" technique. About half the retarded subjects were organically impaired. Analyses of the mean scores for the conservation tasks indicated that conservation of mass was easiest and conservation of volume most difficult for both the retarded and the nonretarded group, providing some support for the notion that the order of emergence of these types of conservation in groups of both average and below-average IQ is as follows: mass, weight, then volume.

In another article McManis (1969e) reported his assessment of conservation and transitivity of weight (clay) and length (sticks) in what appears to be the same sample used in his 1969d experiment. The study was designed to test the hypothesis, derived from Kooistra (1964), that for any given property (e.g., weight) conservation will appear developmentally earlier than transitivity. The results supported this hypothesis for both weight and length in retarded and nonretarded subjects. McManis (1970) then explored the relations among conservation, seriation, and transitivity (of length) within groups of 80 institutionalized mentally retarded persons (IQs of 46–72) and 80 nonretarded elementary school children (IQs of 85–116). Among both retarded and nonretarded children who showed discrepant performance on the conservation and seriation tasks, nearly all showed conservation without seriation. Among both retarded and nonretarded children who showed discrepant performance on the seriation and transitivity tasks, nearly all showed seriation without transitivity. These findings indicate that seriation falls developmentally between conservation and transitivity (at least with respect to the property of length, as measured in this experiment) for both retarded and nonretarded persons.

In a fourth article based on the same sample used in two of the preceding studies (McManis, 1969d, 1969e), McManis (1969b) tested Piaget's (1952) view that there are three hierarchically ordered stages in the development of quantitative comparison processes. In the first stage, children are said to consider only uncoordinated perceptual relations of gross qualitative equality or difference; in the second stage, intensive quantity, children are said to compare quantities by seriating them along more than one dimension (e.g., width and height) simultaneously; in the third stage, extensive quantities, children are said to be capable of overruling apparent differences between two equal quantities by imposing equal units of measurement upon them. McManis tested his young subjects' performance of these three types of comparison, using sticks, colored water, and beads. The analysis of scores on these tasks indicated that for both the retarded and the nonretarded group, gross comparisons were the simplest (they were passed by nearly all subjects in both groups) and extensive comparisons were the most difficult. These findings are consistent with the view that for children at both IQ levels the developmental order is as follows: gross, intensive, and extensive quantity (the order posited by Piaget). One other study (McManis, 1969a) should be mentioned in this connection. McManis's tests of quantity com-

parison were given to 140 institutionalized mentally retarded persons, who were divided into equal groups representing different IQ levels (IQs of 30–49 and of 50–69). The procedure and analyses were similar to those employed in the preceding study (McManis, 1969b). The results indicated, as in the preceding experiment, that comparison of gross quantities was easiest and comparison of extensive quantities was most difficult, regardless of the IQ level of the subjects.

Three other comparative studies were designed to address the problem of order of events across the types of conservation. Gruen and Vore (1972) assessed conservation of number (poker ships), continuous quantity (water), and weight (clay rolled into various shapes) in familial retarded (IQs of 55–80) and nonretarded (IQs of 90–120) public school pupils. Both retarded and nonretarded groups were divided into three subgroups of MA: 5, 7, and 9 years. Evidence on developmental ordering was in the form of mean scores for the three types of conservation, analyzed within each MA level. In one set of analyses, conservation judgments alone (i.e., disregarding the subjects' verbal explanations) constituted the dependent variable. With this criterion, performance of nearly all subjects at MA 9 was correct; for the other two MA levels, both retarded and nonretarded subjects tended to score significantly better on the number task than on the quantity or weight tasks. For nonretarded children at MA 7, however, the differences were not significant. The quantity and weight tasks did not differ significantly in difficulty for retarded and nonretarded children.

In a second set of analyses by Gruen and Vore, the dependent measure was conservation judgment in combination with the subject's explanation for that judgment. Using this criterion, there was no significant task effect at the MA 5 level. At the MA 7 level both retarded and nonretarded subjects did somewhat better at the number task than at the quantity and weight tasks, but the differences were only significant for the retarded subjects. At the MA 9 level both retarded and nonretarded subjects scored significantly higher on the number than on the weight task and significantly higher on the quantity than on the weight task. Thus, although order-of-difficulty patterns were similar for retarded and nonretarded subjects, with use of judgments alone and judgments plus explanations, task differences tended to be statistically significant among the retarded more often than among the nonretarded. In attempting to account for this trend, Gruen and Vore (1972) made an interesting point:

> McManis (1969[e]) has suggested that there is a transitional period (MAs of 7–10) in which the various concrete operations are obtained and that retarded children progress through this period more slowly than do normal children. If this is true, it would be expected that the performance of normal children on various conservation tasks would vary less from task to task than that of retarded children. This also suggests that retarded children may be ideal subjects for investigating the transition process from preoperational to concrete-operational thinking. (p. 156)

We conclude this section of the review by discussing two conservation studies by Achenbach. Building on the work of Charlesworth (1969) and Mermelstein and Shulman (1967), Achenbach (1973) inferred children's identity concepts with respect to color, number, length, and continuous quantity from their surprise reactions to contrived changes in those properties. For example, to test number identity concepts two toy Indians were placed in a box, and when the bottom was opened three Indians dropped out. Among nonretarded subjects (M IQ = 116), there were significantly more frequent surprise reactions to a change in color than to changes in the three quantitative properties, a finding consistent with Piaget's view (see Piaget & Voyat, 1968) that children develop identity concepts for qualitative properties such as color prior to the emergence of identity concepts for quantitative properties such as number, length, and continuous quantity. Surprise reactions to change in color and number were virtually identical in 45 familial and 16 Down syndrome retarded subjects (M IQ = 47). The frequencies of surprise reactions to changes of the three different quantitative properties for both retarded and nonretarded subjects are consistent with the findings of others (including Gruen & Vore, 1972) that successful performance on conventional length and number conservation tasks is simpler than, and thus presumably developmentally prior to,

success on conventional continuous quantity tasks. Thus, once again we see fairly strong support for the view that the sequence of developmental events for the retarded child is similar to that for the nonretarded child.

A different type of similarity is illustrated by Achenbach's (1969) study of nonretarded public school children (IQs of 94–168) and nonorganically impaired retarded children (IQs of 31–78) from public schools and institutions. He assessed children's conservation concepts with respect to length, area, and volume[3] (employing 4 tasks for each of the three properties) by using optical illusions to create discrepancies between the actual and apparent sizes of various stimuli. To test conservation of length, for example, the experimenter presented each child with a barbell illusion in which a small metal rod that fit into a groove that just touched the inner edges of two circles was placed into another groove that passed through two circles and touched their outer edges. The effect was to make the rod appear longer in the second position than in the first. Subjects were then asked whether the rod would fit into the original groove. An important feature of the study, for our purposes, is that the 12 tasks were designed to be free of intellectual demands in the areas of additivity, numeration, conservation of equivalence, or complex verbal expression—dimensions along which more traditional conservation tasks often vary. This made it possible to test the contention of Braine and Shanks (1965a, 1965b) that the attainment of conservation with respect to the various properties would be parallel if the performance criteria used were standard across the types of conservation. Consistent with the Braine–Shanks view, Achenbach (1969) found a "total absence of evidence for a horizontal *décalage*" (p. 677) in the three types of conservation, for both the retarded group and the nonretarded group. In both groups, there were neither consistent nor significant differences in success rates for length, area, and volume tasks. This finding, together with the reasoning of Braine and Shanks, suggests that some of the order-of-difficulty and scalogram-type evidence reviewed in the preceding paragraphs may be more indicative of differences in the specific requirements of the contrived tasks employed than of actual differences in the order of emergence of the various types of conservation.

Concepts of time. A number of studies have addressed the similar sequence hypothesis in content areas other than conservation. In one study on the concept of time, Lovell and Slater (1960) interviewed 50 educationally subnormal children (IQs not reported) aged 8, 9, 10, 11, and 15 years and 50 "average to above average" children aged 5–9 years. The interview included tasks (some were Piaget's) designed to measure concepts of simultaneity and equality of synchronous intervals (e.g., asking the child to judge whether two dolls traveling at different rates but starting and stopping at the same time actually traveled for the same amount of time). There were also tasks involving chronological ordering of events and children's concepts of age and interior time. Little in the way of statistical analysis was reported, but nonetheless, Lovell and Slater concluded that the understanding of these five concepts of time follows roughly the same sequence in retarded as in normal children, although the stages in understanding are reached some years later by retarded children.

Concepts of space. Two studies that included a retarded sample were specifically concerned with spatial concepts. In one of these, Houssiadas and Brown (1967) sampled 40 institutionalized, mentally retarded Australians (M IQ = 55) who showed no evidence of mongolism or other specific defects and who ranged in age from 8–15 years. These subjects were presented with two perspective-taking tasks (one pictorial and one using manipulation of actual objects) in which they were asked to identify their own perspective on a perceptual array as well as the perspective of another person seated at a different position. Although no statistical analyses were reported, the pattern of passes and failures on the different items was consistent with the view that retarded individuals pass through the stages identified by Piaget; that is, first there is difficulty in identifying one's own perspective

[3] The volume conservation task used by Achenbach (1969) actually tapped what Piaget (has) called conservation of *continuous quantity*.

and that of another, second there is only difficulty in identifying how a perceptual array might look from another position, and third the individual is able to "coordinate perspectives," identifying not only his own perspective but that of another person as well. Summing up, Houssiadas and Brown (1967) concluded, "It is clear that the pattern of predominant responses follows the same sequence suggested by Piaget, whose data were derived from normal children" (p. 213).

In a more fine-grained analysis of spatial concepts, Woodward (1962) tested the same institutionalized retarded group of 50 adults and 44 children used in the study of number concepts described earlier (Woodward, 1961). In this sample, 50% of the adults and 61% of the children showed some type of organic impairment. The spatial tasks included measures of the ability to reproduce a spatial order under varying degrees of transformation (e.g., reproducing a circular array of beads on a horizontal rod). Using a similar procedure with nonretarded children, Piaget and Inhelder (1956) identified seven stages through which their children passed as they improved on the tasks. Woodward (1962) constructed a table of scale types to assess the comparability of her results with those of Piaget and Inhelder. Although no scalogram statistics were calculated, the great majority of Woodward's subjects fit scale types consistent with the developmental sequence posited by the Genevans.

A second task employed by Woodward involved drawing copies of 21 geometric figures used by Piaget and Inhelder. The compatibility of subjects' scores on these tasks with a four-stage sequence advanced by Piaget and Inhelder (1956) was demonstrated by the fact that "subjects classified by the features of a given stage showed the features of the lower stages in most cases" (Woodward, 1962, p. 31). However, once again no scalogram statistics were reported, and 5 of the 14 performance criteria by which stage assignments were made were outside the appropriate difficulty level for at least some of the subjects. The third task employed by Woodward was a "reference points" problem in which adult subjects only were presented with drawings of a bottle tilted at various angles and said to be about one-fourth full of water. The subjects' task was to pencil in the portion of the bottle occupied by the water. The performance data presented for this task were extremely sketchy, but Woodward indicated that the order of difficulty of the tasks was the same as that found by Piaget and Inhelder. In her overview of her findings bearing on what we have called the similar sequence hypothesis, Woodward (1962) concluded that for her retarded subjects, "The sequence suggested by Piaget and Inhelder [for nonretarded children] was confirmed for all three spatial concepts that were investigated" (p. 35).

Relative thinking. In investigating the "logic of relations" in children, Piaget (1928) used a "brothers and sisters" problem and a "right and left" problem. In the former problem, children's understanding of the relation between being and having a sibling was explored by asking such questions as "George has three brothers, Paul, Henry, and Charles. How many brothers has Paul? How many brothers are there in this family?" In the right and left problem, children were instructed, "Show me your right hand, your left. Show me my right hand, my left" Lane and Kinder (1939) used these two Piagetian problems with 50 institutionalized retarded individuals of unspecified etiology who were grouped, for purposes of data analysis, into four different IQ levels: 38, 51, 64, and 77. Instead of scalogram statistics, relative levels of the questions were reported for each IQ group. These data indicated that the rank ordering of difficulty for the 11 questions was similar across the different IQ levels—a parallelism consistent with the similar sequence hypothesis.

Moral judgment. Abel (1941) investigated moral judgment in 74 institutionalized "subnormal adolescent white girls" (aged 15–21 years; IQs unspecified). Subjects were questioned about seven brief stories concerning immanent punishment (the inevitability of punishment following a misdeed), retributive justice (punishment orientation, particularly of the "eye for an eye" variety), and judgments of the gravity of a misdeed (using information on consequences of the deed and intent of the transgressor). Mirroring previous findings with

nonretarded individuals (Lerner, 1937, 1938), Abel's findings were that with increasing maturity (defined in terms of MA) subjects gave nonsignificantly greater weight to intent and less weight to consequences in judging the gravity of a misdeed and were significantly less likely to consistently advocate retributive punishment. Unlike the nonretarded persons in at least some research, Abel's more mature subjects (MAs of 9–11 years) did not show any less pronounced a belief in immanent punishment than did her less mature subjects (MAs of 6–8 years). In fact, about 82% of both groups showed such a belief, which Abel (1941) attributed to the "constraining" institutional environment "that controls the girls with threats of immanent punishment" (p. 386). Except for this one anomaly, the Abel data are consistent with the similar sequence hypothesis.

Studies of multiple concepts. We conclude this section on cross-sectional research with a discussion of studies that have assessed concepts in more than one conceptual domain. DeVries (1970, 1973a, 1973b, 1974) assessed a variety of Piagetian concepts in bright (M IQ $\approx$ 130), average (M IQ $\approx$ 105), and retarded (M IQ $\approx$ 72); etiologies not reported) children, all enrolled in public schools. The tasks included the brothers and sisters and right and left problems described earlier; tests of generic and sex identity and of conservation of mass, number, length, and liquid; interviews on magic and dream concepts; object sorting and class inclusion problems; and a guessing game ("Which hand has the penny?") designed to reveal the level of children's role-taking skills. Of all the tasks used, data from the guessing game task were presented in the most complete manner (see DeVries, 1970). Using an independent sample of 64 high-IQ children, DeVries (1970) classified behavior on the guessing game with respect to 10 characteristics (e.g., does not always hide penny in the same hand). These characteristics formed a highly reproducible Guttman-type scale with a reproducibility of .95 and an index of consistency of .66. The scale was then used with the bright, average, and retarded samples and checked against Kohlberg's (1969) criteria for developmental sequentiality, namely, (a) mean scale scores should increase with age, (b) success on each individual scale item should increase with age, and (c) the sequence of items should be justifiable with a logical rationale based on Piagetian theory. DeVries (1973b) maintained that her scale met the third criterion, and her data (DeVries, 1970) indicate that the first two criteria were met within the bright, average, and retarded groups separately. Similar analyses were carried out with respect to the other 14 Piagetian tasks, with a Guttman scale constructed for each. Within the average and retarded groups each scale met Green's (1956) criterion of an index of consistency greater than .50, and the lowest coefficient of reproducibility was .94 (DeVries, Note 1). DeVries (1973b) indicated that all of the Kohlberg criteria for sequentiality "were applied to each ability group (i.e., bright, average, and retarded subjects) separately, and the order of scale items was the same for each ability group on all tasks" (p. 3).

Stearns and Borkowski (1969) investigated conservation of continuous quantity (water) and discontinuous quantity (blocks and marbles) as well as horizontal–vertical space perception in institutionalized retarded individuals (IQs unspecified) ranging in age from $7\frac{1}{2}$–27 years. Consistent with Piaget's (e.g., 1964) view (and supporting findings; see Elkind, 1961) that conservation of continuous quantity is more difficult and emerges developmentally later than conservation of discontinuous quantity, Stearns and Borkowski found performance on their test of the former concept to be significantly poorer than performance on their two tests of the latter concept. Scores were also highly similar for the tests of horizontal and vertical frames of reference; this finding is consistent with Piaget's (see Piaget & Inhelder, 1956) view that concepts of the vertical and of the horizontal are acquired at the same time.

Finally, we turn to two studies by Lovell and his colleagues (Lovell, Healey, & Rowland, 1962; Lovell, Mitchell, & Everett, 1962). The studies reported few relevant statistical analyses, but the diversity of concepts examined makes them worthy of brief attention. In the Lovell, Mitchell, and Everett study, groups of nonretarded and educationally subnormal individuals (no IQs reported)

were divided into separate age groups. The skills investigated included additive and multiplicative classification (of objects and pictures differing in multiple dimensions), seriation, multiplication of asymmetrical transitive relations, hierarchical classification, class inclusion, and visual and tactile classification. For all tasks the tabled data indicated a general improvement in performance with increasing age level for both normal and subnormal groups (no significance tests reported).

In the study by Lovell, Healey, and Rowland, the subjects were again groups of nonretarded and educationally subnormal persons from special schools (IQs unreported) who were divided into separate age groups. The groups were presented with 12 of the tasks used by Piaget, Inhelder, and Szeminska (1960) to study the child's geometric concepts. Within normal and subnormal groups separately, correlation coefficients were calculated that related Piagetian stage levels on the 12 tasks to subjects' age levels. Of the 24 coefficients, 23 were significant at the .01 level. In both Lovell et al. studies, the details of subject selection, experimental procedure, and statistical analyses are so skimpy that the findings must be regarded as only suggestive. Nonetheless, although they are not by any means definitive, the data are in harmony with the similar sequence hypothesis.

Summary of the Cross-Sectional and Order-of-Difficulty Evidence on the Similar Sequence Hypothesis

Thus far we have reviewed 28 studies in which cross-sectional and order-of-difficulty evidence is reported in ways that have some bearing on the similar sequence hypothesis. The degree of retardation involved in the samples ranged from profound to mild; the retarded persons sampled ranged in age from childhood to adulthood, were both institutionalized and noninstitutionalized, and included both cultural–familial cases and individuals with diverse organic and emotional disorders. The nonretarded contrast groups, when employed, ranged from slightly below average to extremely high in IQ. The studies reported also varied widely in their experimental methodology and in their methods of data analysis. Despite this great diversity in methodology and in sample characteristics, the data reviewed show rather consistent support for the similar sequence hypothesis, both in its conservative form, which applies only to nonretarded and familial retarded persons (Zigler, 1969), and in its broader form, in which universality of developmental sequence is held to be independent of individual subject characteristics such as organic impairment (see Kohlberg, 1969).

There were very few exceptions to this generalization: (a) Woodward (1959) found that among her profoundly retarded subjects 1 sensorimotor task out of 11 proved to be slightly more difficult than another task that Piaget designated as being one sensorimotor stage higher. (b) Among Rogers's (1977) profoundly retarded subjects, a causality task designed to be at Stage 6 proved to be easier than one of the tasks at Stage 5, whereas individual spatiality tasks within each stage had to be combined to yield a highly reproducible scale. (c) In Achenbach's (1973) familial and Down's syndrome retarded sample, surprise reactions indicative of color and number identity occurred with equal frequency, whereas in his nonretarded sample, color surprise was significantly more frequent than number surprise. (d) Abel (1941) found no significant decline with increasing MA in belief in immanent punishment in her institutionalized female retarded sample, a finding that differs from some earlier research with nonretarded children.

These four instances of no support for the similar sequence hypothesis are exceedingly minor. They may have resulted, as the authors of these studies generally suggested from idiosyncratic (or misinterpreted) properties of the tasks selected or from other measurement errors, and in some cases (e.g., Abel, 1941) they may reflect the suppressive influence of an idiosyncratic environment that merely delays the shift from one level of reasoning to another. Furthermore, in each of the four studies, the findings of no support were outnumbered by findings supporting the similar sequence hypothesis.

While noting the strong level of support that the reviewed evidence has yielded for the similar sequence hypothesis, we must also note that such cross-sectional and level-of-

difficulty evidence, even at its best, can support only indirect inference regarding the actual process of development. Considerably more direct and potent inference is possible when an investigator observes the same individuals at more than one point during the course of development, that is, in longitudinal fashion. Such research is often expensive and complex, and consequently it is relatively rare, particularly with mentally retarded persons. However, three longitudinal studies have some bearing on the similar sequence hypothesis. We now turn to these.

Longitudinal Evidence

Development in the sensorimotor period. One longitudinal investigation was designed by Wohlhueter and Sindberg (1975) as an extension of Woodward's (1959) cross-sectional study of sensorimotor development in profoundly retarded persons (described earlier in this article). These investigators conducted monthly assessments of institutionalized 1–6-year-old profoundly, severely, and moderately retarded children (no IQs reported). The Piagetian object concept tasks used by Décarie (1965) were employed for 1 to $1\frac{1}{2}$ years or until a child performed at the highest of the 10 substage levels for 2 consecutive monthly sessions. Of the principal sample of 49 children, 20 had progressed to the highest substage level by the end of the study; of the remaining 29, 10 showed a generally monotonic increase, and 9 seemed to be at a plateau, with object concept levels the same for most of the 12 or more sessions. Thus, 39 of the 49 subjects showed patterns harmonious with the pattern of object concept stages posited by Piaget (1955) and found by subsequent investigators using nonretarded samples. For the remaining 10 subjects, however, there was a variable developmental pattern in which substage levels appeared to rise and fall from session to session, ranging over as many as 3 or 4 substages during the 12 or more sessions.

In an effort to determine what characteristics might distinguish this group of atypical subjects, Wohlhueter and Sindberg, (1975) examined medical histories and clinical findings for their sample; the distinguishing feature of the variable group was that the majority of subjects "were found to have EEG abnormalities, especially dysrhythmias or a history of seizures" (p. 516). This finding raises at least two possible interpretations with respect to the variable developmental pattern: (a) that individuals with brain anomalies associated with electroencephalogram (EEG) abnormalities may show atypical sequences of development with respect to the object concept and (b) that behavioral and attentional abnormalities in individuals with anomalous EEG patterns make accurate assessment of object concept substages difficult.

One other unusual pattern was noted by the investigators, namely, some children seemed to bypass or skip over some of the substages. This apparent skipping phenomenon has been noted in research with nonretarded children as well (see Uzgiris, Note 2). Although this is an interesting phenomenon, it is difficult to know how often skipped substages may actually have been traversed by the children in the intervals between experimental sessions. Furthermore, both the skipping phenomenon and the variability phenomenon might be better understood if we were able to rule out specific method effects, as could have been done if a nonretarded sample had been included in this study for comparative purposes. Nonetheless, the Wohlhueter–Sindberg investigation at least raises significant questions about the validity of the similar sequence hypothesis with respect to certain substages in the development of the object concept.

A longitudinal study reported by Cicchetti and Sroufe (1976), however, yields strong support for the similar sequence hypothesis within the sensorimotor period. The study focused on the relation between cognitive and affective development in home-reared Down's syndrome infants during the period from 4–18 months of age. Sroufe and his colleagues (e.g., Sroufe & Wunsch, 1972) earlier demonstrated that among normal infants there is a developmental progression from mirth response to auditory and tactile stimulation that is physically intense or vigorous (e.g., tickling the baby's chin or saying "Boom!") to mirth responses to social and visual stimulation that is progressively more subtle and complex (e.g., the sight of mother sucking on baby's bottle). At monthly intervals the infants sampled by Cicchetti and Sroufe were pre-

sented with 15 auditory and tactile items of the intense or vigorous type and 15 social and visual items of the more subtle and complex type. As in the earlier research with normal infants, the Down's syndrome infants laughed earliest in response to the auditory and tactile items and latest in response to "the more cognitively complicated social and visual items" (Cicchetti & Sroufe, 1976, p. 923). The responses of the Down's infants, of course, came months later (in CA) than did the corresponding responses of normal infants. Smile responses, a more sensitive index of positive affect in the Down's syndrome sample, showed the same pattern and revealed even more clearly than laughter responses the developmental decline in positive affect aroused by simpler auditory and tactile items as the infants matured beyond 13 months. This inverted-U-shaped developmental pattern also resembles earlier findings with normal infants. In stressing the similarity of their findings with Down's infants to those with normal infants, Cicchetti and Sroufe (1976) pointed out that the laughter items were similarly ordered for both groups, "category by category and, in the main, item by item" (p. 923). Finally, to assess the merits of their claim that the affective responses they measured were closely related to cognitive development, Cicchetti and Sroufe calculated correlations of indices of affective expression (e.g., earliest laugh, total amount of smiling to all items, etc.) with the Bayley mental and motors scales and the Uzgiris–Hunt object permanence and operational causality scales. All 44 correlations were statistically significant.

Therefore, the Cicchetti–Sroufe investigation, unlike the Wohlhueter–Sindberg study, provides uniform support for the liberal version of the similar sequence hypothesis in which the hypothesis is applied to all retarded children regardless of etiology. The Cicchetti–Sroufe research deserves special attention because of its unusually careful methology and its emphasis on the integrity of the developing infant. The research demonstrates a thoughtful means of assessing the Piagetian hypothesis that affective development and cognitive development are interdependent. In addition, it may help to point the way to tests of the similar sequence hypothesis in behavioral domains other than cognitive development as it has traditionally been construed.

Research on the preoperational–concrete operational transition—the Temple longitudinal study. The third longitudinal study reviewed here is by far the broadest in scope. In this ongoing investigation, Stephens and her colleagues (Stephens, 1974; Stephens, Mahaney, & McLaughlin, 1972; Stephens et al., 1974) have conducted biennial assessments of the performance of retarded and nonretarded persons on a variety of Piagetian tasks. The sample included 75 retarded subjects (IQs of 50–75) from special education classes and 75 nonretarded subjects (IQs of 90–110) from the same Philadelphia schools. In the first wave of testing, the age range in both subject groups was 6–18 years. Results from the first two waves of testing have now been reported and are discussed here within two content categories:

Moral judgment. The Temple battery included 11 measures designed to assess three aspects of moral judgment: (a) the relative weight assigned to intent versus consequences in judging the seriousness of a misdeed, (b) awareness of the injustice of punishing an entire group for the acts of only one or a few members, and (c) the ability to judge the relative fairness of various types of punishment including retributive and reciprocal justice. To determine whether judgment along these three dimensions follows the same developmental course in the retarded as in the nonretarded, Mahaney and Stephens (1974) examined changes in scores on the 11 component measures over the 2-year period from the first to the second wave of testing. They found that on 1 of the intent-versus-consequences measures the retarded group showed a nonsignificant decline in score (i.e., they made slightly less mature moral judgments according to the scoring criteria adopted by the authors) and that on 1 of the group punishment items nonretarded subjects showed a nonsignificant decline. On the 9 remaining items the direction of change was the same for both the retarded subjects and the nonretarded subjects; this similarity extended to 2 items on which both groups showed significant de-

clines in score, raising questions about the scoring of these particular items.

Inhelder (1943/1968), in a study described earlier, referred to certain "oscillations" in the reasoning of the retarded. It is of some interest to note that Mahaney and Stephens (who was the translator of the Inhelder book) reported oscillations, that is, instances when "the improvement which occurred in one area of moral judgment was not maintained when opinions were solicited on another, but similar, situation" (Mahaney & Stephens, p. 137), in both retarded and nonretarded subjects. There is some indication that such oscillations may have been somewhat more frequent in the retarded group. One line of evidence that suggests this possibility comes from Mahaney and Stephens's analysis of change scores for three separate age levels within both the retarded and the nonretarded samples. Of the 29 change scores reported for the nonretarded groups, 20 were increases (11 significant), 7 were decreases (3 significant), and 2 showed no change. Among the retarded subjects, 17 change scores were increases (7 significant), 11 were decreases (4 significant), and 1 involved no change. However, when retarded comparison groups were staggered in order to broaden the age difference involved in age group comparisons, for example, by comparing Phase 1 6–10-year-olds with Phase 2 12–16-year-olds, the only items showing a decrease with age were the two that showed a decrease in nonretarded subjects as well. Overall, the report by Mahaney and Stephens (1974) suggests that although growth in moral judgment concepts among the retarded may be "torporific and sporadic" (p. 141), the direction of development is the same for both retarded and nonretarded persons.[4]

Conservation, classification, symbolic imagery, and formal operations in the Temple study. The Temple investigation (Stephens, 1974; Stephens et al., 1972; Stephens et al., 1974) also included 29 measures of cognitive development across four broad conceptual domains: (a) conservation (of substance, length, weight, continuous quantity, and volume, as well as term-to-term correspondence), (b) logic classification (class inclusion and class intersection, and relative thinking measured by the brothers–sisters and right–left tests), (c) operativity and symbolic imagery (tests involving imagined rotations of objects through space, transferring from two to three dimensions, and changing one's perspective on a stimulus), and (d) combinatory logic (Piaget & Inhelder's 1956, combination of liquids task). Explanations given by subjects on each of the 29 items were scored on a 9-point scale that took into account, among other things, the degree to which the subject wavered between a correct and an incorrect answer and the degree to which reversibility was shown. Stephens and McLaughlin (1974) reported on changes in scores on the 29 measures over the 2-year period separating the two waves of testing. They found that the nonretarded group showed improvement on all 29 measures, with 25 statistically significant; the retarded group also improved on all 29 measures, with 26 statistically significant. This finding indicates that the direction of development on these Piagetian reasoning tasks was similar in the retarded and nonretarded groups. In another report following the second wave of the Temple study (Stephens et al., 1972), the Piagetian reasoning tasks were rank ordered with respect to the MAs at which 50% of the subjects (in the retarded and nonretarded groups separately) made correct responses. As Stephens et al. indicated, the order of difficulty for both subject groups was generally consistent with previous findings that conservation of substance precedes conservation of weight

[4] Whether this apparent similarity in the direction of development is actually a function of an invariant, stagelike progression is thus far an open question because of the questionable nature of the measures themselves. In a critique of this portion of the Temple longitudinal study, Kohlberg (1974) maintained that Piagetian moral judgment measures used in the Temple study do not even warrant detailed longitudinal analysis, because

> Piaget himself does not consider that his moral judgment measures yield genuine stages, nor do they pair up with his logical stages in ways compatible with his current thinking about cognitive stages . . . . Empirical research confirms the fact that Piaget's moral stage measures do not meet the criteria of structural stages which his logical stages do meet. (p. 142)

This being the case, it is appropriate to be cautious about what one concludes with respect to the moral judgment portion of the Temple investigation.

and that conservation of weight precedes conservation of volume. In addition, although the ranks were rather crude because MA levels were listed in whole years and because not even the most ardent Piagetian would expect all 29 items to form an orderly developmental scale, it is interesting to note that our own calculations yielded a Spearman rho of .634 between the rank order given for the retarded group and that given for the nonretarded group.

The preceding data are consistent with the similar sequence hypothesis, as far as they have been taken, but they could be taken considerably further. With the one exception mentioned in the preceding paragraph, there has been no apparent effort thus far by the Temple investigators to check their findings against specific developmental stage sequences such as the horizontal decalages reported for nonretarded subjects in previous Piagetian research (Kohlberg & DeVries, 1971; Nassafat, 1963; Siegelman & Block, 1969; Smedslund, 1964; Uzgiris, 1968). Moreover, there has been a persistent inclination to report data in terms of group means, rather than in terms of the number of individuals (retarded and nonretarded) who show specific developmental patterns. This latter type of analysis is the unique province of longitudinal research and can only be approximated indirectly by scaling procedures in research of the nonlongitudinal variety. There is some indication (see Stephens, 1974) that efforts to profile individual performance changes over time and to cross-validate specific vertical and horizontal decalages will be forthcoming from the Temple investigators. Such efforts are needed if the investigators are to fully capitalize on the power of their longitudinal design.

Status of Evidence on the Similar Sequence Hypothesis

Only 1 of the 3 longitudinal studies reviewed —the Wohlhueter and Sindberg (1975) investigation of object concept substages—produced findings inconsistent with the similar sequence hypothesis. In that investigation a distinct subgroup of 10 (out of 49) children showed apparent atypical developmental sequences, and most of these children showed anomalous EEG patterns. This finding may indicate that brain wave anomalies can be associated with atypical developmental patterns. Alternatively, the EEG abnormalities may simply have been associated with attentional and other deficits that interfered with accurate assessment of substage levels in children whose actual development was consistent with the similar sequence hypothesis. The latter interpretation has special credence in the area of object concept assessment, in which procedures demand that the subject sustain attention to an object long enough to seek after it once it has been removed from the preceptual field. Of the 28 nonlongitudinal studies reviewed, only 4 contained a finding inconsistent with the similar sequence hypothesis, and in each of these studies the inconsistent finding was relatively minor and was alone among a number of findings supporting the hypothesis. Furthermore, the questions raised generally concerned rather fine-grained steps or substages within horizontal decalages on which studies of nonretarded subjects alone have not always agreed.

These facts, plus the measurement problems inherent in these experimental procedures, make the degree of consistency in the findings of these 31 studies rather striking. Positive findings have now been reported in conceptual areas that include sensorimotor spatial concepts, object permanence, causality, imitation, affective responding, identity and equivalence conservation (of many properties), seriation, transitivity, moral reasoning, comparison processes (or gross, intensive, and extensive quantities), time, space, relative thinking, role taking, mental imagery, geometric concepts, and classification and class inclusion. For the 31 studies spanning this list of conceptual areas, the great preponderance of the evidence is consistent with the hypothesis that retarded and nonretarded persons traverse the same stages of development in the same order, differing only in the rate at which they progress and in the ultimate ceiling they attain. The hypothesis seems to be generally supported in studies of retarded individuals, regardless of etiology, with the possible exception of individuals suffering from pronounced EEG abnormalities.

Quality of the Evidence

Having said this, we believe it is important to comment on the quality of the available evidence and to offer suggestions for improving it. Cross-sectional data relevant to the similar sequence hypothesis have most often been presented in ways that provide only the weakest inferential power. A table displaying the percentage of subjects at each age level who pass each Piagetian item can yield only a rather faint glimmer of developmental sequence compared with the information generated when each child is classified with respect to specific pass–fail patterns, that is, with respect to response scale types. When such a scaling analysis is combined with calculation of scalogram summary statistics (e.g., Green, 1956; Guttman, 1950), the potential power of the nonlongitudinal design is more fully utilized.

Similarly, the bulk of the longitudinal data we found was presented only in terms of mean difference between experimental groups or changes in group means or percentages over time. As Hunt (1974) has noted, reporting only group summary statistics at Time 1 and Time 2 can mask the fact that some individuals progressed while others regressed over time. Thus, although it is useful to know that the Time 1 and Time 2 means differed in the same direction for retarded and nonretarded groups, such information is no substitute for an analysis of the number of individuals in each group showing specific developmental patterns over time. In both longitudinal and nonlongitudinal research aimed at testing the similar sequence hypothesis, it makes little sense to invest the time and energy necessary to gather potentially relevant data and then analyze the data in ways that fail to capitalize on their full potential.

Suggestions Toward Improved Research

These problems, and others to which we referred earlier in the text, suggest three principles that if widely adopted would substantially improve the quality of evidence on the similar sequence hypothesis.

Structuring Direct Comparisons

A problem with many of the studies reviewed in earlier sections is that their samples included only mentally retarded subjects. In those few instances in which findings of these studies disagree with findings of other studies sampling only nonretarded subjects, the discrepancies are difficult to interpret. This is because of uncertainty over whether the discrepancies reflect actual process differences between the retarded and the nonretarded or whether the differences in experimental methodology across studies are responsible. An obvious way to prevent such difficulties is to expose retarded and nonretarded children within a similar cognitive developmental range to precisely the same procedure by including both groups in the same study. To fail to do so is to risk uninterpretable findings.

Attending to Etiology

The marked heterogeneity of many of the mentally retarded samples described earlier suggests a somewhat opportunistic approach to subject selection or perhaps an approach in which the etiology of retardation is simply not regarded as an important factor. Yet, theoretical considerations discussed early in this article (see also, Weisz, 1976; Zigler, 1969, 1971) point to the need to give special attention to familial retarded children as opposed to those suffering from organic impairment or genetic disorder. Furthermore, Wohlhueter and Sindberg's (1975) report of atypical development in a group of children with a high incidence of EEG anomalies suggests the potential importance of efforts to identify developmentally distinct subgroups within the nonfamilial population. Their analysis illustrates that subgroup analyses can be useful even when they are post hoc.

Promoting Uniformity

Finally, there is a clear need for increased uniformity across studies in the kinds of statistical analyses carried out and in the way statistics are reported. Toward this end, we suggest that every cross-sectional study addressing the similar sequence hypothesis should yield data bearing on the following threefold question:

1. Within the retarded and nonretarded groups do the task items form the same scale, and does this scale show high reproducibility

(à la Guttman, 1950) and a high index of consistency (à la Green, 1956)?

2. Do mean scale scores increase with level of cognitive maturity in each separate subject group (see Kohlberg, 1969)?

3. Do mean levels of success on each individual item increase with levels of cognitive maturity in each separate subject group (see Kohlberg, 1969)?

In longitudinal research, Questions 2 and 3 should also be asked in a way that only longitudinal investigation permits: Over the period spanned by the longitudinal study, in what percentage of individual subjects (from retarded and nonretarded groups) do scale scores and individual item scores (a) increase either smoothly or monotonically, (b) remain stable throughout, and (c) show at least some declines.

These recommended questions are designed to promote greater uniformity, and thus greater comparability, among studies addressing the similar sequence hypothesis. In opposition to such uniformity, one might argue that the degree of consistency in the findings of the numerous studies reviewed here is all the more impressive precisely because of the methodological diversity of the studies. There is some truth to this argument, but only in those cases in which findings support the similar sequence hypothesis. However, we have argued that even in those cases apparent group similarities in developmental sequence may result from a failure to ask the most probing questions of one's data. It seems clear from our review that evidence from the 31 studies currently available offers rather consistent support for the similar sequence hypothesis; it also seems likely that the best evidence has yet to be gathered.

Reference Notes

1. DeVries, R. Personal communication, February 1, 1978.
2. Uzgiris, I. C. *Some antecedents of the object concept.* Paper presented at the meeting of the Eastern Psychological Association, Philadelphia, April 1969.

References

Abel, T. M. Moral judgments among subnormals. *Journal of Abnormal and Social Psychology*, 1941, *36*, 378–392.

Achenbach, T. M. Conservation of illusion-distorted identity: Its relation to MA and CA in normals and retardates. *Child Development*, 1969, *40*, 663–679.

Achenbach, T. M. Surprise and memory as indices of concrete operational development. *Psychological Reports*, 1973, *33*, 47–57.

Balla, D., Styfco, S. J., & Zigler, E. Use of the opposition concept and outerdirectedness in intellectually average, familial retarded, and organically retarded children. *American Journal of Mental Deficiency*, 1971, *75*, 663–680.

Balla, D., & Zigler, E. Discrimination and switching learning in normal, familial retarded, and organic retarded children. *Journal of Abnormal and Social Psychology*, 1964, *61*, 664–669.

Benton, A. L. Behavioral indices of brain injury in school children. *Child Development*, 1962, *33*, 199–208.

Braine, M. D. S., & Shanks, B. The conservation of a shape property and a proposal about the origins of the conservations. *Canadian Journal of Psychology*, 1965, *19*, 197–207. (a)

Braine, M. D. S., & Shanks, B. The development of conservation of size. *Journal of Verbal Learning and Verbal Behavior*, 1965, *4*, 227–242. (b)

Buck-Morss, S. Socio-economic bias in Piaget's theory and its implications for cross-culture studies. *Human Development*, 1975, *18*, 35–49.

Charlesworth, W. R. The role of surprise in cognitive development. In D. Elkind & J. H. Flavell (Eds.), *Studies in cognitive development: Essays in honor of Jean Piaget.* New York: Oxford University Press, 1969.

Cicchetti, D., & Sroufe, L. A. The relationship between affective and cognitive development in Down's syndrome infants. *Child Development*, 1976, *47*, 920–929.

Cruikshank, W. M. *The brain-injured child in home, school, and community.* Syracuse, N.Y.: Syracuse University Press, 1967.

Décarie, T. G. *Intelligence and affectivity in early childhood.* New York: International Universities Press, 1965.

DeVries, R. The development of role-taking as reflected by behavior of bright, average, and retarded children in a social guessing game. *Child Development*, 1970, *41*, 759–770.

DeVries, R. *Performance on Piaget-type tasks of high-IQ, average-IQ, and low-IQ children.* Paper presented at the meeting of the Society for Research in Child Development, Philadelphia, Pa., 1973. (ERIC Document Reproduction Service No. ED 086 374/PS 007 129) (a)

DeVries, R. *The two intelligences of bright, average, and retarded children.* Paper presented at the meeting of the Society for Research in Child Development, Philadelphia, Pa., 1973. (ERIC Document Reproduction Service No. ED 079/102/SE 016 419) (b)

DeVries, R. Relationships among Piagetian, IQ, and achievement assessments. *Child Development*, 1974, *45*, 746–756.

Elkind, D. The development of quantitative thinking. *Journal of Genetic Psychology*, 1961, *98*, 36–46.

Elkind, D. Piaget's conservation problems. *Child Development*, 1967, *38*, 15–27.

Elkind, D., Koegler, R. R., Go, E., & Van Doorninck, W. Effects of perceptual training on unmatched samples of brain-injured and familial retarded children. *Journal of Abnormal Pscychology*, 1965, *70*, 107–110.

Ellis, N. R. A behavioral research strategy in mental retardation: Defense and critique. *American Journal of Mental Deficiency*, 1969, *73*, 557–566.

Green, B. F. A method for scalogram analysis using summary statistics. *Psychometrika*, 1956, *21*, 79–88.

Gruen, G. E., & Vore, D. A. Development of conservation in normal and retarded children. *Developmental Psychology*, 1972, *6*, 146–157.

Guttman, L. The basis of scalogram analysis. In S. A. Stouffer et al. (Eds.), *Measurement and prediction* (Vol. 4). Princeton, N.J.: Princeton University Press, 1950.

Harter, S., Brown, L., & Zigler, E. The discrimination learning of normal and retarded children as a function of penalty conditions and etiology of the retarded. *Child Development*, 1971, *42*, 517–536.

Houssiadas, L., & Brown, L. B. The coordination of perspectives by mentally retarded children. *Journal of Genetic Psychology*, 1967, *110*, 211–215.

Hunt, J. McV. Discussion: Developmental gains in reasoning. *American Journal of Mental Deficiency*, 1974, *79*, 127–133.

Inhelder, B. *The diagnosis of reasoning in the mentally retarded*. New York: Day, 1968. (Originally published, 1943)

Jordan, V. B. *Cognitive development among retardates: Reanalysis of Inhelder's data*. Paper presented at the meeting of the Society for Research in Child Development, Denver, Colo., 1976. (ERIC Document Reproduction Service No. EB 121 035/EC 082 713)

Klauss, S. D., & Green, M. B. Conservation in trainable mentally retarded children. *Training School Bulletin*, 1972, *69*, 108–114.

Kohlberg, L. Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), *Handbook of socialization theory and research*. Chicago: Rand McNally, 1969.

Kohlberg, L. From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel (Ed.), *Cognitive development and epistemology*. New York: Academic Press, 1971.

Kohlberg, L. Discussion: Developmental gains in moral judgment. *American Journal of Mental Deficiency*, 1974, *79*, 142–146.

Kohlberg, L., & DeVries, R. Relations between Piagetian and psychometric assessments of intelligence. In C. Lavatelli (Ed.), *The natural curriculum*. Urbana: University of Illinois Press, 1971.

Kooistra, W. H. Developmental trends in the attainment of conservation, transitivity, and relativism in the thinking of children: A replication and extension of Piaget's ontogenetic formulations (Doctoral dissertation, Wayne State University, 1963). *Dissertation Abstracts*, 1964, *25*, 2032. (University Microfilms No. 64-9538)

Lane, E. B., & Kinder, E. F. Relativism in the thinking of subnormal subjects as measured by certain of Piaget's tests. *Journal of Genetic Psychology*, 1939, *54*, 107–118.

Lerner, E. *Constraint areas and the moral judgment of children*. Memasha, Wis.: Banta, 1937.

Lerner, E. Observations sur le raisonnement moral de l'enfant. *Cahiers pédagogiques experimentaux et psychologiques de l'enfant* (Vol. No. 11). Geneva: Palais Wilson, 1938.

Lister, C. The development of ESN children's understanding of conservation in a range of attribute situations. *British Journal of Educational Psychology*, 1972, *42*, 14–22.

Lovell, K., Healey, D., & Rowland, A. D. Growth of some geometric concepts. *Child Development*, 1962, *33*, 751–767.

Lovell, K., Mitchell, B., & Everett, I. R. An experimental study of the growth of some logical structures. *British Journal of Psychology*, 1962, *53*, 175–188.

Lovell, K., & Slater, A. The growth of the concept of time: A comparative study. *Child Psychology and Psychiatry*, 1960, *1*, 179–190.

Mahaney, E. J., & Stephens, B. Two-year gains in moral judgment by retarded and nonretarded persons. *American Journal of Mental Deficiency*, 1974, *79*, 134–141.

Mannix, J. B. The number concepts of a group of E. S. N. children. *British Journal of Educational Psychology*, 1960, *30*, 180–181.

Marchi, J. U. Comparison of selected Piagetian tasks with the Wechsler Intelligence Scale for Children as measures of mental retardation (Doctoral dissertation, University of California, Berkeley, 1970). *Dissertation Abstracts International*, 1971, *31*, 6442A. (University Microfilms No. 71-51, 833).

Mermelstein, E., & Shulman, L. S. Lack of formal schooling and the acquisition of conservation. *Child Development*, 1967, *38*, 39–52.

McManis, D. Comparison of gross, intensive, and extensive quantities by retardates. *Journal of Genetic Psychology*, 1969, *115*, 229–236. (a)

McManis, D. Comparisons of gross, intensive, and extensive quantities by normals and retardates. *Child Development*, 1969, *30*, 237–244. (b)

McManis, D. Conservation of identity and equivalence of quantity by retardates. *Journal of Genetic Psychology*, 1969, *115*, 63–69. (c)

McManis, D. Conservation of mass, weight, and volume by normal and retarded children. *American Journal of Mental Deficiency*, 1969, *73*, 762–767. (d)

McManis, D. Conservation and transitivity of weight and length by normals and retardates. *Developmental Psychology*, 1969, *1*, 373–382. (e)

McManis, D. Conservation, seriation, and transitivity performance by retarded and average individuals. *American Journal of Mental Deficiency*, 1970, *74*, 784–791.

Milgram, N. A. Cognition and language in mental retardation: Distinctions and implications. In D. K. Routh (Ed.), *The experimental psychology of mental retardation*. Chicago: Aldine, 1973.

Nassafat, M. *Etude quantitative sur l'èvolution des opèrations intellectuelles: Le passage des opèrations concrètes aux opèrations formelles*. Neuchâtel, Switzerland: Delachaux et Niestlé, 1963.

Papalia, D., & Hooper, F. A developmental comparison of identity and equivalence. *Journal of Experimental Child Psychology*, 1971, *12*, 347–361.
Piaget, J. *Judgment and reasoning in the child*. New York: Harcourt, Brace, 1928.
Piaget, J. *The child's conception of number*. New York: Humanities Press, 1952.
Piaget, J. *The origins of intelligence in the child*. London: Routledge & Kegan Paul, 1953.
Piaget, J. *The construction of reality in the child*. London: Routledge & Kegan Paul, 1955.
Piaget, J. The general problem of the psychobiological development of the child. *Discussions on Child Development*, 1956, *4*, 3–27.
Piaget, J. *Play, dreams, and imitation in childhood*. New York: Norton, 1962.
Piaget, J. Cognitive development in children: The Piaget papers. In R. E. Ripple & V. N. Rockcastle (Eds.), *Piaget rediscovered: A report of the Conference on Cognitive Studies and Curriculum Development*. Ithaca, N.Y.: Cornell University, School of Education, 1964.
Piaget, J. Nécessité et signification des recherches comparatives en psychologie génétique. *International Journal of Psychology*, 1966, *1*, 3–13.
Piaget, J. Preface. In B. Inhelder, *The diagnosis of reasoning in the mentally retarded*. New York: Chandler, 1968.
Piaget, J. Piaget's theory. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology* (3rd ed.). New York: Wiley, 1970.
Piaget, J. *The psychology of intelligence*. Totowa, N.J.: Littlefield, Adams, 1972.
Piaget, J., & Inhelder, B. *Le développement des quantités chez l'enfant*. Neuchâtel, Switzerland: Delachaux et Niestlé, 1941.
Piaget, J., & Inhelder, B. *The child's conception of space*. New York: Humanities Press, 1956.
Piaget, J., Inhelder, B., & Szeminska, A. *The child's conception of geometry*. New York: Routledge & Kegan Paul, 1960.
Piaget, J., & Voyat, G. Recherche sur l'identité d'un corps en développement et sur celle du mouvement transitif. In J. Piaget, H. Sinclair, & Vinh Bang (Eds.), *Epistémologie et psychologie de l'identité: Etude d'épistémologie génétique* (Vol. 24). Paris: Presses Universitaires de France, 1968.
Popper, K. R. *The logic of scientific discovery*. New York: Basic Books, 1959.
Reitan, R. M. Psychological assessment of deficits associated with brain lesions in subjects. In J. L. Khanna (Ed.), *Brain damage and mental retardation: A psychological evaluation*. Springfield, Ill.: Charles C Thomas, 1973.
Rogers, S. J. Characteristics of the cognitive development of profoundly retarded children. *Child Development*, 1977, *48*, 837–843.
Roodin, P. A., Sullivan, L., & Rybash, J. M. Effects of a memory aid on three types of conservation in institutionalized retarded children. *Journal of Genetic Psychology*, 1976, *129*, 253–259.
Siegelman, E., & Block, J. Two scalable sets of Piagetian tasks. *Child Development*, 1969, *40*, 951–956.
Simpson, E. L. Moral development research: A case study of scientific culture bias. *Human Development*, 1974, *17*, 81–106.
Singh, N. N., & Stott, G. The conservation of number in mental retardates. *Australian Journal of Mental Retardation*, 1975, *3*, 215–221.
Smedslund, J. Concrete reasoning: A study of intellectual development. *Monographs of the Society for Research in Child Development*, 1964, *29*(2, Serial No. 93).
Sroufe, L. A., & Wunsch, J. P. The development of laughter in the first year of life. *Child Development*, 1972, *43*, 1326–1344.
Stearns, K., & Borkowski, J. G. The development of conservation and horizontal-vertical space perception in mental retardation. *American Journal of Mental Deficiency*, 1969, *73*, 785–790.
Stephens, B. Symposium: Developmental gains in the reasoning, moral judgment, and moral conduct of retarded and nonretarded persons. *American Journal of Mental Deficiency*, 1974, *79*, 113–115.
Stephens, B., Mahaney, E. J., & McLaughlin, J. A. Mental ages for achievement of Piagetian reasoning assessments. *Education and Training of the Mentally Retarded*, 1972, 7, 124–128.
Stephens, B., & McLaughlin, J. A. Two-year gains in reasoning by retarded and nonretarded persons. *American Journal of Mental Deficiency*, 1974, *79*, 116–126.
Stephens, B., et al. Symposium: Developmental gains in the reasoning, moral judgment, and moral conduct of retarded and nonretarded persons. *American Journal of Mental Deficiency*, 1974, *79*, 113–161.
Uzgiris, I. Situational generality of conservation. In I. Sigel & F. Hooper (Eds.), *Logical thinking in children: Research based on Piaget's theory*. New York: Holt, Rinehart & Winston, 1968.
Weisz, J. R. Studying cognitive development in retarded and nonretarded groups: The role of theory. *American Journal of Mental Deficiency*, 1976, *81*, 235–239.
Weisz, J. R. A follow-up developmental study of hypothesis behavior among mentally retarded and nonretarded children. *Journal of Experimental Child Psychology*, 1977, *34*, 108–122.
Weisz, J. R. Transcontextual validity in developmental research. *Child Development*, 1978, *49*, 1–12.
Weisz, J. R., & Achenbach, T. M. Effects of IQ and mental age on hypothesis behavior in normal and retarded children. *Developmental Psychology*, 1975, *11*, 304–310.
Weisz, J. R., & Zigler, E. Developmental versus difference theories of mental retardation: The developmental evidence. In E. Zigler & D. Balla (Eds.), *Developmental and difference theories of mental retardation*. Hillsdale, N.J.: Erlbaum, in press.
Wilton, K. M., & Boersma, F. J. Conservation research with the mentally retarded. *International Review of Research in Mental Retardation*, 1974, *7*, 113–144.
Wohlhueter, M. J., & Sindberg, R. M. Longitudinal development of object permanence in mentally retarded children: An exploratory study. *American Journal of Mental Deficiency*, 1975, *79*, 513–518.
Woodward, M. The behavior of idiots interpreted by

Piaget's theory of sensorimotor development. *British Jounral of Educational Psychology*, 1959, *29*, 60–71.

Woodward, M. Concepts of number of the mentally subnormal studied by Piaget's method. *Journal of Child Psychology and Psychiatry*, 1961, *2*, 249–259.

Woodward, M. Concepts of space in the mentally subnormal studied by Piaget's method. *British Journal of Social and Clinical Psychology*, 1962, *1*, 25–37.

Woodward, M. The application of Piaget's theory to research in mental deficiency. In N. R. Ellis (Ed.), *Handbook of mental deficiency*. New York: McGraw-Hill, 1963.

Zigler, E. Developmental versus difference theories of mental retardation and the problem of motivation. *American Journal of Mental Deficiency*, 1969, *73*, 536–556.

Zigler, E. The retarded child as a whole person. In H. E. Adams & W. K. Boardman (Eds.), *Advances in experimental clinical psychology*. New York: Pergamon Press, 1971.

Received April 3, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 852-875

Stability of Aggressive Reaction Patterns in Males: A Review

Dan Olweus
University of Bergen, Bergen, Norway

Considered in the review are 16 studies on the stability of aggressive behavior and reaction patterns. There is great variation among the studies in sample composition, in definition of variables, in method of data collection, and in the ages and intervals studied. Generally, the size of a (disattenuated) stability coefficient tends to decrease linearly as the interval between the two times of measurement ($T_2 - T_1$) increases. Furthermore, the degree of stability can be broadly described as a positive linear function of the interval covered and the subject's age at the time of first measurement, expressed in the age ratio T_1/T_2. The degree of stability that exists in the area of aggression was found to be quite substantial; it was, in fact, not much lower than the stability typically found in the domain of intelligence testing. Marked individual differences in habitual aggression level manifest themselves early in life, certainly by the age of 3. It was generally concluded that (a) the degree of longitudinal consistency in aggressive behavior patterns is much greater than has been maintained by proponents of a behavioral specificity position, and (b) important determinants of the observed longitudinal consistency are to be found in relatively stable, individual-differentiating reaction tendencies or motive systems (personality variables) within individuals.

What has become known as the consistency issue has been the subject of recent lively discussions in the professional literature of psychology. This issue, which has often been presented in an overly simplified way, concerns a whole complex of problems, and it is obvious that different forms of consistency can be conceived of (cf. Olweus, 1974). In the recent discussions, the main emphasis has been on the question of cross-situational consistency (see Endler & Magnusson, 1976), which primarily concerns the extent to which individuals in a group retain their relative positions on a certain dimension or characteristic across various situations, conditions, or sources of data at approximately the same point in time. This issue seems to be far from settled, and in many ways the debate has been confusing, characterized by strong emotional reactions, stereotyped presentations of different theoretical positions, and methodological mistakes (for critical analyses, see, e.g., Block, 1977; Epstein, 1977; Golding, 1975; Olweus, 1977b). In the general debate, the issue of longitudinal consistency or stability—which concerns the extent to which individuals in a group retain their relative positions on a certain dimension or characteristic (or similar dimensions or characteristics) for measurements at different periods of time—has attracted much less attention. Mischel (1968, 1969) however, dealt with the issue at some length and concluded that there is generally little longitudinal (as well as cross-situational) consistency in noncognitive personality dimensions (with the exception of self-descrip-

---

This article is based on invited addresses given at the International Conference on Psychological Issues in Changing Aggression, Warsaw, Poland, July 9–14, 1976, and at the International Congress of Psychology, Paris, France, July 18–25, 1976. The author's research reported in this article was supported by grants from the Norwegian Research Council for Science and the Humanities and the Swedish Council for Social Research.

I want to thank Jack Block for valuable comments on an earlier draft of this article.

Requests for reprints should be sent to Dan Olweus, Department of Psychology, Box 25, N-5014 University of Bergen, Bergen, Norway.

tions on trait dimensions). But the empirical material presented by Mischel in support of his position appears both meagre and selective (see also, Block, 1977). Considering the general importance of the question of longitudinal consistency for personality psychology, it seems valuable to take a closer look at the empirical evidence available. Here this is done for one particular area of research, aggression, which no doubt represents an important behavior system in psychology. Aggression was included among the areas reviewed by Mischel.

Furthermore, the general usefulness of personality concepts involving relatively enduring internal factors or properties of individuals has recently been questioned. For instance, Krasner and Ullmann (1973) tried to demonstrate that the concept of personality is superfluous as a descriptive or explanatory term. They wrote, "The more we know about antecedent, current, and consequent conditions, the less likely we are to use the concept of personality" (p. 489). Similarly, but from a partly different point of view, Shweder (1975) attempted to show that a personality concept in its individual difference sense is of little value and relevance. Also in consideration of these and similar views, it is important to examine the evidence on the stability or continuity of aggressive behavior over time. Such data no doubt will help to shed light on the question of the general utility of assuming a relatively enduring personality system or at least certain more stable subsystems within the individual.

The general purpose of this article is twofold. One goal is to get a picture of the degree of stability obtaining in the area of aggression, as manifested in longitudinal studies of aggressive reaction patterns in males (with the exclusion of self-reports; see below). The main conclusions on this point are presented under the headings of General Description of Results and More Specific, Descriptive Conclusions. A second purpose is to interpret these stability data with a particular view to the possibility of considering them as partly reflecting relatively stable, individual-differentiating reaction tendencies within individuals. Furthermore, the empirical stability data and the suggested interpretations are briefly related to the views expressed by proponents of a behavioral specificity position such as Mischel (1968, 1969) and by critics of personality concepts such as Krasner and Ullmann (1973). This goal is pursued in the sections entitled Interpretation of the Stability Data and Conclusions.

With regard to the second purpose, it should be pointed out that I am generally sympathetic to the view that personality variables in terms of relatively stable, individual-differentiating reaction tendencies may be important (but by no means the sole) determinants of an individual's aggressive behavior (see, e.g., Olweus, 1969, 1973b, 1978). This point of departure is not likely to have influenced the present data gathering and data analysis procedure. However, it seems fair to state at the outset that the analysis in the interpretative section, which leaves somewhat greater room for judgment, has been approached from this perspective.

Coverage of Review

This article is a shortened version of a review presented elsewhere (Olweus, Note 1). In the more complete report relatively detailed descriptions of all studies considered are given. Because of space limitations, such descriptions are provided here for only three reports—one with preschool subjects, the second with subjects of school age (two studies in one article), and the third with adults. However, summary data on method of observation, age and number of subjects, interval beteen measurements, and so on are presented in Table 1 for all studies reviewed.[1]

The focus of the present review is on longitudinal studies of aggressive behavior and reaction patterns, as observed or inferred by individuals other than the subjects themselves. It is important to note that the present review thus does not include studies centering on the stability of self-descriptions

[1] Readers who are interested in detailed information on the studies not described in the present article can obtain a copy of the more complete report by writing to the author.

Table 1
Longitudinal Studies of Aggression[a]

Study	Method of data collection or integration	N	Reliability at T_1	Reliability at T_2	Age at T_1	Interval in years $(T_2 - T_1)$	Age ratio (T_1/T_2)	Raw correlation (uncorrected)	Correlation corrected for attenuation
		Subjects below age 6 at T_1							
Patterson, Littman, & Bricker (1967)	Direct observation	36[b]	.80[e]	.80	3.5	.50	.88	.72	.90
Kohn & Rosman (1972)	Teacher ratings	70	.77	.73	4.0	.50	.89	.56	.74
Jersild & Markey (1935)	Direct observation	24[c]	.80	.80	3.0	.75	.80	.70	.88
Emmerich (1966)	Teacher ratings	53[d]	.85[e]	.85	3.5	.83	.81	.65	.76
Martin (1964)	Direct observation	53[d]	.80[e]	.80	3.5	.83	.81	.52	.65
Block, Block, & Harrington (1974)	Teacher ratings	41	.86	.74	3.5	1.00	.78	.70	.88
Kohn & Rosman (1972)	Teacher ratings	250	.77	.77	4.0	1.00	.80	.53	.69
Kohn & Rosman (1972)	Teacher ratings	250	.77	.83	4.0	1.50	.73	.48	.60
Kohn & Rosman (1973)	Teacher ratings	271	.70	.70	5.0	1.50	.77	.51	.73
Kagan & Moss (1962)	Clinical ratings	36	.70	.90[e]	5.0	18.00	.22	.22	.26
Kagan & Moss (1962)	Clinical ratings	36	.60	.90[e]	2.0	21.00	.09	.29	.36

Table 1 (*continued*)

Study	Method of data collection or integration	N	Reliability at T_1	Reliability at T_2	Age at T_1	Interval in years $(T_2 - T_1)$	Age ratio (T_1/T_2)	Raw correlation (uncorrected)	Correlation corrected for attenuation
		Subjects above age 6 at T_1							
Wiggins & Winder (1961)	Peer nominations	163	.87	.87	9.0	1.00	.90	.56	.65
Wiggins & Winder (1961)	Peer nominations	176	.81	.81	10.0	1.00	.91	.54	.67
Olweus (1977a)	Peer ratings	85	.81	.81	13.0	1.00	.93	.80	.98
Block (1971)	Clinical ratings	84	.66	.74	12.0	3.00	.80	.54	.69
Olweus (1977a)	Peer ratings	201	.85	.87	13.0	3.00	.81	.68	.79
Farrington (1978)	Teacher ratings	410	.80[e]	.80	9.0	4.00	.69	.51	.64
Eron, Huesmann, Lefkowitz, & Walder (1972)	Peer nominations	71	.90	.82	8.0	5.00	.62	.48	.56
Eron, Huesmann, Lefkowitz, & Walder (1972)	Peer nominations	71	.82	.85	13.0	5.00	.72	.65	.78
Eron, Huesmann, Lefkowitz, & Walder (1972)	Peer nominations	211	.90	.85	8.0	10.00	.44	.38	.44
Kagan & Moss (1962)	Clinical ratings	36	.85	.90[e]	13.0	10.00	.57	.56	.67
Kagan & Moss (1962)	Clinical ratings	36	.85	.90[e]	9.0	14.00	.39	.40	.48
Tuddenham (1959)	Clinical ratings	32	.75	.74	18.0	14.00	.56	.68	.91
Block (1971)	Clinical ratings	84	.74	.78	15.0	18.00	.46	.44	.53

Note. T_1 = time of first measurement; T_2 = time of second measurement.
[a] Samples on which stability correlations were based generally included males only, except where indicated by a superscript.
[b] Subjects were 18 males and 18 females.
[c] Subjects were male and female.
[d] Subjects were 29 males and 24 females.
[e] See Olweus (Note 1) for details.

on personality questionnaires, self-ratings on trait scales, and similar self-report devices. The stability of this kind of data has often been considered quite substantial (although the self-descriptions are often considered tenuously related to the actual behavior to which they refer; see, e.g., Mischel, 1968, 1969). Some additional comments on the stated guidelines for inclusion in the review are in order.

In most cases the term *aggressive* was used by the original investigator in his or her specification of the variable under study. By and large, these specifications seem to be in accordance with the definition of an aggressive response, given elsewhere (Olweus, 1973b) as

> any act or behavior that involves, might involve, and/or to some extent can be considered as aiming at, the infliction of injury or discomfort; also manifestations of inner reactions such as feelings or thoughts that can be considered to have such an aim are regarded as aggressive responses. (p. 270)

Although most studies reviewed concern data on aggressive interpersonal behavior, some variables deal primarily with aggressive reactivity (e.g., "over-reactive to minor frustrations; irritable"; Block, 1971). Occasionally an author has assigned a variable to the aggressive behavior system that is not included here. This is the case with variables such as competitiveness, dominance, and repression of aggressive thoughts, which are more indirect manifestations of aggressive (and other) tendencies or may be assumed to reflect conflict over, or inhibitions against, aggressive tendencies rather than the aggressive tendencies themselves.

As mentioned, the stated guidelines also imply that studies concerned with the longitudinal stability of inventory responses, self-ratings, and so on are not considered in the review. In one of the included studies (Block, 1971), however, such self-report data were used in combination with other sets of data as a basis for clinical ratings. Longitudinal studies in which projective instruments constituted the only source of data (if such studies exist in the aggressive-motive area) are also excluded from consideration. The exclusion of stability data derived from the latter two data sources makes the material for review somewhat more homogeneous. This, however, does not preclude a substantial variation in a number of respects among the studies considered, as becomes evident.

A further criterion for inclusion in the review is that the degree of stability of the data has been expressed in the form of a correlation coefficient or that such a coefficient can be derived from the reported data in a meaningful way. Accordingly, studies in which only significance tests between discrete groups have been presented are excluded. Furthermore, since the main focus is on the degree of stability/change in behavior rather than on the direction of possible changes, no attention is paid to differences in mean levels between different periods of time; besides, such data are often not available in the reports.

It has been possible to locate 16 studies comprising 16 independent male or mixed samples of subjects for which stability data have been collected. (The 4 studies on mixed nursery school groups are considered here to consist of 1 sample each.) These 16 studies have been described in 14 publications. One of the studies is English (Farrington, 1978), two are based on Swedish samples (Olweus, 1977a), and the rest used American subjects.[2] In some studies, however, the same sample was employed for the determination of several stability correlations covering different periods of time. A total of 24 stability coefficients are available.

In a limited number of studies (six) there was also an independent female sample from which stability data were collected. The female samples are not considered in the present context (except for a brief mention in Footnote 6).

The discussion of the criteria for exclusion of studies from this review may give the impression that there exist a large number of studies on the stability of aggressive behavior and reactions that have not been

[2] It should be mentioned that the search of literature was mainly restricted to books and professional journals written in English.

considered. It is important to emphasize that this is definitely not the case.

Correction for Attenuation

As is well-known, a stability correlation between two sets of measurements on a particular variable is systematically lowered (attenuated) as the result of errors of measurement. Accordingly, the correlation between true or error-free scores on the same variable is higher than that between fallible scores. If the reliabilities of the two sets of measurements are known, the formula for the correction for attenuation can be used to compute the disattenuated correlation, that is, an estimate of the correlation between corresponding true or perfectly reliable scores. The attenuation formula is as follows: $r_{tt} = r_{xy}/(r_{xx} \cdot r_{yy})^{\frac{1}{2}}$, where r_{tt} is the correlation between true scores of x and y, r_{xy} is the obtained correlation between x and y, and r_{xx} and r_{yy} are the reliability coefficients of x and y (see, e.g., Lord & Novick, 1968). It should be noted that the sampling error of coefficients corrected for attenuation is greater than that of uncorrected coefficients of the same size (Thouless, 1939). Accordingly, it is reasonable to expect disattenuated coefficients to occasionally exceed unity as a function of sampling fluctuations, particularly if the size of the sample is relatively small. Correlation coefficients corrected for attenuation should always be considered only approximate in character.

In spite of the latter circumstances, disattenuated coefficients can be very useful and under certain conditions are the most appropriate measures to employ. If for instance, a number of stability correlations are to be compared and they are based on data of varying reliability, attenuation correction will make the coefficients more directly comparable. Furthermore, if the researcher's primary interest is in the relationship between the true rather than the obtained scores, for example, in the stability of the underlying function(s) in contrast with the actual predictive power of the fallible measurements, a disattenuated coefficient is the correct measure to use (see, e.g., Block, 1963, 1971; Lord & Novick, 1968; Thouless, 1939). For both these reasons, attenuation-corrected coefficients are reported in the following review in addition to the uncorrected raw correlations. Most of the theoretical analyses and conclusions are based on disattenuated coefficients.

Relatively little has been written in the psychometric literature with respect to the type of reliability coefficient to be used in the denominator of the correction formula. However, the general principle to follow seems clear and is stated as follows by Lord and Novick (1968):

> If we are to use the correction for attenuation precisely, we must use it only in conjunction with an experimental design that assures that essentially no more or less error variation is introduced into the estimate of r_{xx} and r_{yy} than is introduced into the estimate of r_{xy} [notation changed in accordance with usage here]. (p. 138)

Since many sources contribute to variations among observations (Lord & Novick, 1968, p. 140), the application of the stated principle to a concrete situation is not always simple and straightforward. In general, the most serious error in the present context would occur if the disattenuated coefficients were overcorrected or inflated as a result of too low, deflated reliability estimates. In some of the studies reported here, several reliability estimates were available for possible use in the attenuation formula. In addition, some studies presented either incomplete reliability information or none at all, and so "best guesses" about the reliability estimates had to be made. In such decisions care was taken to settle on reliability estimates that seemed too high rather than too low, in order to avoid the risk of having inflated, disattenuated coefficients. Exact information about the reliability of the measures was not available in some studies, of course, resulting in a somewhat greater uncertainty as to the correct size of the disattenuated coefficients in these studies. It should be noted, however, that the guesses made were of an informed character, as reliability information from similar studies was used. The number of stability coefficients for which best guesses about reliability estimates were made amounted to 5 out of 24 (Em-

merich, 1966; Farrington, 1978; Jersild & Markey, 1935; Martin, 1964; Patterson, Littman, & Bricker, 1967). In all probability, the potential error introduced by this procedure is not great and in any event, because of the strategy adopted, is not likely to have resulted in inflated coefficients.

Aspects of Studies to Be Considered in the Review

To arrive at relatively specific conclusions, it was considered necessary to present the studies under review in some detail. A description of the samples and of the procedures employed in collecting the data is provided, including an at least approximate indication of the definition of the variables studied or of possible interest. In the few cases when a variable of potential relevance has been excluded from the review, the reasons for the decision are given (in the unabridged report, Olweus, Note 1). The ages of the subjects at the time of the first (T_1) and later (T_2, T_3, etc.) measurements are also reported. In some situations, however, a particular problem arises with respect to the subjects' age. This occurs when the material on which the judges' assessments were based does not refer to a specific point in time but covers a period of several years. To relate the size of the stability coefficients to the interval between the two times of measurements (or more precisely, the times to which the measurements refer), the "exact" age of the subjects must be determined. In establishing a rule for the determination of the exact age in such cases the following line of reasoning was applied. It was assumed that a judge who was to assess, for example, archival material or a retrospective interview embracing several years, would have given relatively more weight to information pertaining to the end of the period covered. Accordingly, it seemed reasonable to fix the exact age as 1 year minus the subject's age at the end of the period in question. For example, when the archival material in a particular study referred to the period between ages 3 and 6, the exact age of the subjects was taken to be 5 years. The stated rule was applied in 4 out of 16 studies, or to 9 of the 24 periods for which stability correlations were reported.

Furthermore, the studies reviewed were scrutinized for information regarding environmental changes during the interval from the first to the later time of measurement. Such data are of relevance when it comes to interpreting the stability results obtained. Although information on this point is seldom detailed or individualized, it may give a rough idea of the degree of environmental change characterizing the periods covered.

The reliabilities of the measures or best estimates of such measures are, of course, presented in addition to obtained and attenuation-corrected stability coefficients.

In two of the studies (Block, 1971; Kagan & Moss, 1962) it was natural and possible to compute stability correlations for composite variables, consisting of two components. This was done by means of Spearman's (1913) formula for the correlation of sums (see Olweus, Note 1).

It should be noted that these composites represent aggression variables of greater scope and generality than their components. No doubt it is of considerable theoretical interest to assess the stability of such more generalized variables. They may be of particular value and relevance if the interval separating the two times of measurement is long and there is some uncertainty about the conceptual equivalence of the different variables being used at the different periods of time.

The review briefly presents data, where available, about relationships between the aggression variables studied and information from other, independent modes of measurement. Although incomplete, this information provides some idea of the concurrent, predictive, and construct validity of the data and concepts under study. By this procedure, the reader obtains a better basis for evaluating the adequacy of the conclusions reached. The procedure also furnishes some evidence on the degree of cross-situational consistency, in the sense of the amount of correspondence between aggression data from different, independent sources or modes of measurement.

Description of Selected Studies

Jersild and Markey (1935)

In this early study the behavior of each of 54 children aged 2–4 years (*M* age = 3 years) was recorded during 10 distributed 15-minute periods of free play. The children, 30 boys and 24 girls, were enrolled in three different nursery school groups. Approximately 9 months later 24 of these children were observed again with the same method and during the same number of sessions. The subjects were still divided in three different groups, a total of some 54 children. Relatively marked changes in the peer group composition took place from the first to the second time of observation, particularly for two of the groups (B and C).

The reliability of the observations, as a measure of the adequacy of the sampling of behavior (generalizability across sessions and, to some extent, observers), was determined in several different ways. For the variable of primary interest in the present context, frequency of being the aggressor, the Spearman-Brown corrected coefficients varied between .31 and .90 at the first period of observation, although the majority of them were above .70. At the second period, the two stepped-up coefficients reported were .81 and .55. As a best estimate of the reliability of the observations, a value of .80 was used for both periods of time and was inserted in the formula for attenuation correction of the stability correlation. It can also be mentioned that the percentage of agreement between independent observers of the same behavior sessions was generally high, amounting to 96% in the case of frequency of being the aggressor. Five different observers collected data in this study, although one of them obtained more than half of all the records.

The rank-order correlation (rho) for stability over the 9-month interval was .70 when calculated for all 24 children. When correlations were determined separately for the different groups (*n* varying from 7 to 15), the coefficients were even higher (.71–.88). After correction for attenuation the stability correlation of .70 amounted to .88.

Teacher ratings on a selection of the behavior categories employed were also secured for 35 of the children, permitting a study of the relationship between the behavioral observation variables and independent rating data. Although the ratings and the behavior observations were separated by an interval of several months, the average correlation (for two groups) was .54 for frequency of being the aggressor. The corresponding correlation for the related variable, frequency of physical acts of combat, was .63. If these two variables were combined into a more general composite aggression variable and a similar composite measure were formed for the behavioral observations, the correlation very likely would amount to .75 or more (all the data necessary for carrying out the calculations were not available in the report).

Olweus (1977a)

Two short-term longitudinal studies concerning a 1-year and a 3-year interval, respectively, were conducted by Olweus (1977a) on two samples of Swedish adolescent boys. In both studies the same two 7-point peer-rating scales were used; they concerned unprovoked physical aggression against peers ("He starts fights with other boys at school," abbreviated *start fights*) and verbal aggression against a mildly criticizing teacher ("When a teacher criticizes him, he tends to answer back and protest," abbreviated *verbal protest*). Each boy who served as a rater assessed all the boys in his class by placing cards with the names of his classmates below the points of the scale that referred to different frequencies of occurrence (from *very seldom* to *very often*). The rating procedure was individually administered.

In Study 1 the number of raters in each class was three on both occasions, at Grades 6 and 7. In the second study the number of raters in different classes varied somewhat, the average number being four at Grade 6 and five at Grade 9. In general, the raters were chosen on the basis of random selection from each class. Approximately a third of the raters on the second occasion had also served as raters in Grade 6. To examine if memory effects affected the ratings of iden-

tical raters, a special analysis was conducted in Study 1. This analysis gave no evidence that partial use of the same raters on both occasions inflated stability correlations. Also, the disattenuated coefficients for completely independent rater groups at Grades 6 and 7 were almost identical with the results for rater groups who had one rater in common (see Olweus, 1977a, Footnote 3). Since no memory effects were detected for a 1-year interval, such effects can safely be disregarded in Study 2, which covered a 3-year interval.

The subjects of the first study consisted of 85 boys from 7 classes who were rated at the end of Grade 6, when their median age was 13 years, and also 1 year later. In this study only small changes in the composition of the peer groups took place between Grade 6 and Grade 7. All classes, however, had new teachers at Grade 7. Study 2 comprised 201 boys from 18 classes who were rated at the end of Grade 6 and also 3 years later at the end of Grade 9, when their median age was 16 years. The subjects constituted roughly 75% of the whole population of school boys in the community at these grades. They represented a good deal of variation with respect to socioeconomic factors (relatively representative of greater Stockholm). At Grade 6, the 18 classes comprised a total of 214 boys. Three years later, 13 of these boys has disappeared from these schools and 27 new boys had entered the classes. The new boys constituted roughly 12% of the boys then in the classes at Grade 9. Two classes mentioned later underwent marked changes in the composition of the peer group. Furthermore, all classes had new teachers at Grade 9, and 11 of the classes had moved to other school buildings. A certain amount of environmental change thus occurred for this sample from Grade 6 to Grade 9.

The Spearman-Brown corrected reliability coefficients were estimated as .80 for start fights and as .82 for verbal protest in Study 1. In the second study the corresponding coefficients were .83 and .86, respectively, for average ratings of four raters (Grade 6) and .86 and .88, respectively, for five raters (Grade 9). In Table 1, average reliability estimates for the two variables were given.

The stability correlations in Study 1 were .81 for start fights and .79 for verbal protest. After correction for attenuation these coefficients amounted to 1.01 (rounded to 1.00) and .96, respectively. In Study 2, covering a 3-year interval, the uncorrected stability correlations were .65 and .70, respectively. The disattenuated coefficients were .77 and .81, respectively. It should be noted that to make the ratings from different classes more comparable, the average ratings were converted within each class to standard scores (Z scores). By this procedure differences in mean level and in variability between classes and grades were eliminated. It can also be mentioned that scatterplots of the stability correlations in both studies showed the relationships to be regular and clearly linear in form.

As mentioned, two classes in the second study are of particular interest from a stability/change point of view. In one of these, the original Grade-6 class, consisting of 10 boys, was split into two at the beginning of Grade 8 (5 boys were transferred to another class for unknown reasons). At the second period of rating, the original class had been augmented by 8 new boys (with no previous connections with one another), which thus represented a very marked change in the composition of the class. This change notwithstanding, the stability correlations for the original 5 boys were very high and were, in fact, even higher than the corresponding correlations for the total sample. Also, the transfer of the 5 boys to a new class consisting of 9 boys did not seem to reduce the stability of behavior of the latter: The across-time correlations for the core of 9 boys in this class were for both variables higher than the coefficients for the whole sample (for details, see Olweus, 1977a).

It should also be mentioned that change of school did not seem to appreciably affect the degree of stability over time. There were small and inconsistent differences between the across-time correlations for the 11 classes who moved to other school buildings and the 7 classes who did not move.

As regards relationship to other data, it can be reported that peer ratings on start fights and verbal protest were used as cri-

terion variables for two factorially derived inventory scales of a newly developed multifaceted aggression inventory for boys (Olweus, 1973b; Olweus, Note 2). In several independent samples, substantial correlations have been obtained between these two scales, the Physical Aggression scale and the Verbal Aggression scale, and peer ratings of the overt aggressive behavior. For instance, the average correlation in two samples of boys ($n = 98$ and 86, respectively) between the Verbal Aggression scale and verbal protest was .49 (.63 after correction for attenuation; see Olweus, Note 2, p. 39). Correlations of approximately the same magnitude (.45) were obtained between the Physical Aggression scale and its natural counterpart, start fights (the correlation was .58 after correction for attenuation). Both of the inventory scales obtained clearly higher correlations with their matching than with their nonmatching rating dimensions, thereby giving evidence of discriminant validity. Still higher correlations were found when these two scales were linearly combined to predict the general aggressive behavior dimension: start fights(Z) + verbal protest(Z). The mean of the coefficients for the two samples mentioned amounted to .53,[3] the highest value being .58 (.67 after correction for attenuation).

The peer-rating data in Study 2 were collected within the framework of a large-scale project concerning bully and whipping boy problems in the school that has been described in detail elsewhere (Olweus, 1973a, 1974, 1978). In this context, the form master or form mistress of each class was requested to nominate possible bullies and whipping boys according to specific criteria. When these teacher nominations were related to independent peer ratings, a very convincing picture emerged. The bullies (21 boys from the entire Grade-6 population) were rated as much more aggressive, both physically and verbally, than randomly selected control boys (60 boys) and the whipping boys (21 boys). For start fights and verbal protest, $F(2, 99) = 24.30$, $p < .0001$, $\epsilon = .56$, and $F = 32.39$, $p < .0001$, $\epsilon = .62$, respectively. Essentially the same findings were obtained in two additional, somewhat smaller samples of boys. These results are clearly consistent with theoretical expectations. It can thus be concluded that a substantial degree of correspondence has been demonstrated between peer ratings on one hand and teacher nominations and self-report (inventory) data on the other.

In the article by Olweus (1977a), the rating data (four different variables) were also scrutinized for the possible presence of rater biases, irrelevant method variance, and so on. An adaptation of multimethod–multivariable analysis showed strong evidence for discriminant validity according to all three criteria proposed by Campbell and Fiske (1959). All in all, it was concluded that the results obtained consistently attested to the validity and general adequacy of the ratings employed in these studies.

Block (1971)

The subjects of this study, discussed in *Lives Through Time* (Block, 1971) were 84 adolescent boys and men and 87 girls and women, participants in the well-known Oakland growth (Jones, 1938) and Berkeley guidance (Macfarlane, 1938) longitudinal studies. By means of the Q-sort method the subjects were assessed for three different periods of time: the junior high school years, the senior high school years, and when they were in their middle 30s (the average age was approximately 34 years). For the two adolescent periods extensive archival data such as school grades, comments and ratings by teachers, ratings of social or interview behavior by staff members, performance on intelligence and projective tests, self-reports of areas of agreement or disagreement with parents, and so on were available. From this material, for each subject "case assemblies" were developed separately for the junior high school and senior high school periods. The information collected during

[3] In Olweus (1973b, Table 3, p. 312), the correlation between verbal aggression(Z) + physical aggression(Z) and start fights(Z) + verbal protest(Z) in Sample B was erroneously printed as .42. The correct value is .47.

intensive interviews (an average length of 12 hours) when the subjects were in their mid-30s constituted the third data set. It should be emphasized that these three data sets were completely independent.

The material for a particular subject at a particular period was assessed (as a rule) by three clinical psychologists, each functioning independently. No psychologist evaluated a subject at more than one age, and to minimize the influence of judge biases, the psychologists were assigned to cases in systematically permuted combinations. The judges expressed their characterizations of each subject by means of the California *Q*-set procedure (Block, 1961, 1971). The California *Q* set consists of some 100 items or variables for psychodynamic descriptions of personality that are sorted by the judge into a forced-choice (approximately normal) distribution. In the Block (1971) study the three independent *CQ* sorts for each subject were averaged. Only the two variables most directly concerned with aggressive reactions and behavior are considered in the present context. These variables are Variable 34 ("over-reactive to minor frustrations; irritable") and Variable 62 ("tends to be rebellious and nonconforming"; see Appendix A in Block, 1971; certain other variables, e.g., Variables 38 and 94, are of some, though less direct relevance). In accordance with the previously stated rule, the "exact" age of the subjects is taken to be 12 years for the junior high school (JHS) period, 15 years for the senior high school (SHS) period, and 33 years at the time of adult follow-up.

Some social data and life events of potential significance from a stability/change point of view can be briefly mentioned. These data, however, are not individual and specific but broadly characterize the subjects as a group. The families of the subjects came from predominantly middle and upper classes (mainly from Classes 1–4 on a 6-step scale). The majority of the subjects (81%) experienced intact families, with both the mother and the father present through adolescence. On the whole, the sample appears representative of the stable, relatively prosperous Berkeley community at the time of the study. The social status of the subjects as adults was similar to that of their parents, but the overrepresentation of the two highest social classes was more pronounced in the subject group. By the time of the adult interview, 95% of the subjects had been married and 19% had been divorced. The majority had also become parents, with an average production of 2.5 children. Roughly half the subjects had served in the armed forces during World War II; the other half were adolescents during this period. Although the Depression hit the families of the subjects when the subjects were relatively young (at the time of or before the junior high school period), this crisis is likely to have been of significance for the personality development of certain subjects. Particularly for the older subjects (from the Oakland growth study), the occurrence of the Depression may have reduced the degree of stability of some personality characteristics from the high school years to the time of adult assessment. All in all, these data suggest that "although typicality cannot be claimed for the subjects, they nevertheless have lived recognizable American lives" (Block, 1971, p. 24) and have experienced the adaptational tasks that face most adult persons in connection with marriage, parenthood, and occupational career. In sum, it seems reasonable to assume that in the lives of the subjects under study, there was a good deal of environmental pressure for change during the 20 or so years from high school to the time of the adult follow-up.

For the male sample, the stepped-up interrater reliabilities (Block, 1971, Appendix G) were .49 (JHS), .68 (SHS), and .77 (adult) for Variable 34 and .82 (JHS), .80 (SHS), and .78 (adult) for Variable 62. The raw stability correlations for Variable 34 ("over-reactive . . .") were .45 for the JHS–SHS period (an interval of 3 years) and .29 for the SHS–adult period (an interval of 18 years; the results for the JHS–adult period were not given in Block, 1971). After correction for attenuation, these coefficients became .78 and .40, respectively. For Variable 62 ("tends to be rebellious . . ."), the

stability correlations for the JHS–SHS and SHS–adult periods were .58 and .29, respectively, and after attenuation correction were .72 and .37, respectively.

To calculate the stability correlations for the (unweighted) composite of the standardized variables, $34(Z) + 62(Z)$, some additional information was needed (see Olweus, Note 1). By means of the formula for the correlation of two unweighted composites previously mentioned, the uncorrected stability correlation for the composite variable, $34(Z) + 62(Z)$, was found to be .54 for the JHS–SHS and .44 for the SHS–adult periods. The corresponding disattenuated coefficients were .69 and .53, respectively.

The values for the composite variables are reported in Table 1. The reliabilities in this table for the Block (1971) study are average values for Variable 34 and Variable 62 (e.g., the reliability estimate of .66 for JHS is the average of .49 and .82).

It should be noted that in the present study, great care was taken to secure adequate and valid information by using independent data sets and sophisticated judges working independently in permuted combinations and by a number of additional checks on potential artifacts such as rater biases and stereotypes. The longitudinal data were also analyzed separately for a number of homogeneous personality types derived via inverse factor analysis. For information on these results, the reader should consult Block (1971).

Summary Data on Studies Reviewed

An overview of the studies can be gained from examining Table 1 and Figure 1. In Table 1 the studies are divided into two groups on the basis of the age of the subjects at the time of first measurement (T_1). Within each group the studies are ordered according to the length of the interval separating the two times of measurement ($T_2 - T_1$). Figure 1 presents in diagrammatic form the disattenuated coefficients as a joint function of the subject's age at the time of first measurement and the interval between the two times of measurement. It should be observed that because of space considerations, the reference axes are broken and the units of measurement are different below and above the breaks.

As mentioned, the number of studies with independent samples of subjects was 16. In 4 studies (Block, 1971; Eron, Huesmann, Lefkowitz, & Walder, 1972; Kagan & Moss, 1962; Kohn & Rosman, 1972) the same or partly the same subjects were used in the determination of more than one stability correlation (but for different intervals); a total of 24 stability coefficients are reported in Table 1 and Figure 1. The average raw and disattenuated correlations were .63 and .79, respectively, for the 12 studies with only 1 coefficient per sample, as compared with average values of .55 and .68, respectively, when all 24 stability determinations were used. Although these two materials are not equivalent with regard to the interval covered and so on, the use of several stability coefficients from a limited number of studies does not seem to have resulted in an overrepresentation of high coefficients. In the following discussion the focus is on the characteristics of and the results for the total material.

The average number of subjects on which the stability coefficients were based amounted to 116. The age of the subjects at the time of first measurement varied from 2 to 18 years, and the subjects were followed for intervals varying from half a year to 21 years. The average interval covered was 5.7 years. The highest (average) age of a subject group at the time of follow-up assessment was 33 years.

As is evident from Figure 1, there has been a concentration on short-term longitudinal studies covering intervals up to $1\frac{1}{2}$ years for subjects below school age. Only two stability coefficients are reported for intervals greater than $1\frac{1}{2}$ years, and these coefficients are based on the same, relatively small sample ($N = 36$). For subjects of higher ages, the intervals covered are more evenly distributed.

The methods of data collection or integration were quite varied (see Table 1). In three studies using nursery school groups, the behavior of the subjects was directly

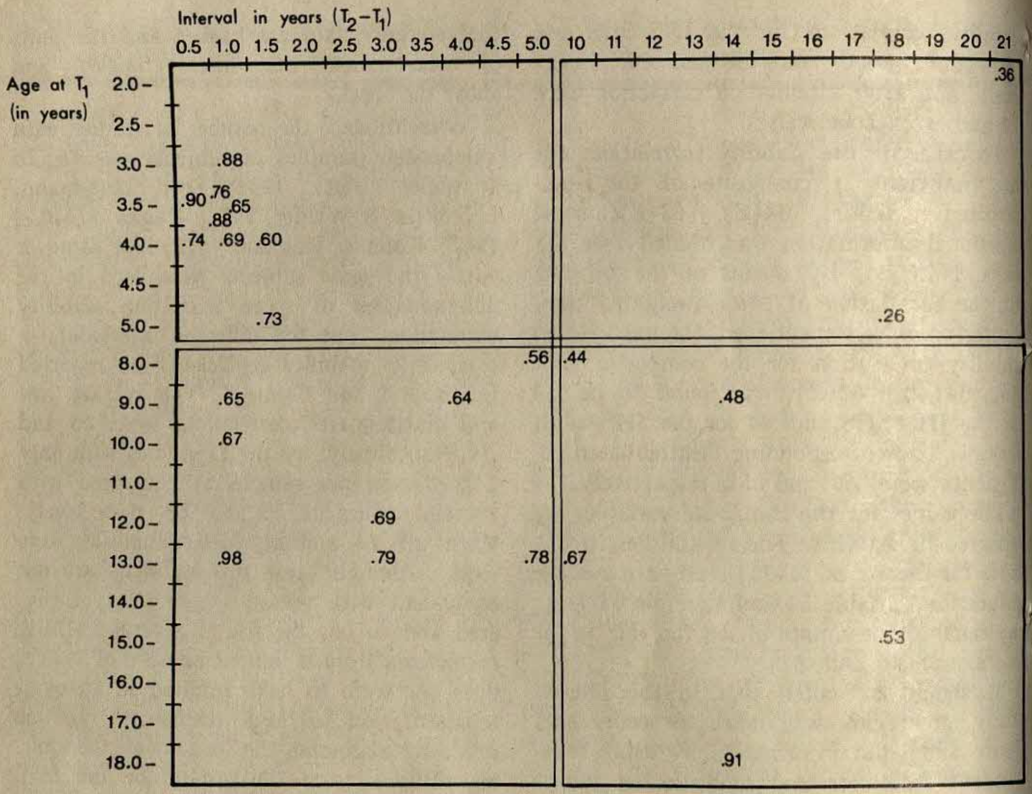

Figure 1. Summary of disattenuated stability correlations for different intervals in years ($T_2 - T_1$) and different ages at time of first measurement (T_1).

observed. Seven stability coefficients were based on teacher ratings, seven on clinical ratings, two on peer ratings, and five on peer nominations.

The variables assessed refer to reactions and behaviors of the subjects in their natural settings, such as nursery school and elementary school. These settings, in which the subjects spent a considerable part of their days at the time of the investigation, can be assumed to represent important sectors of the subjects' lives. The studies by Block (1971), Kagan and Moss (1962), and Tuddenham (1959), which were partly based on archival material collected over several years, provide an unusually broad basis for assessment. In these studies, clinical ratings were used to integrate the rich and diverse material.

Overall, these data suggest a good deal of variation among the studies in a number of important respects. Although a better coverage of certain age periods and intervals would have been desirable, the number of studies is considerable and the average size of the samples quite respectable according to usual standards in psychological research.

Some Methodological Comments

Potential Correlation-Increasing Factors

In four studies using mixed nursery groups (Emmerich, 1966; Jersild & Markey, 1935; Martin, 1964; Patterson et al., 1967), the stability correlations were not computed separately for boys and girls. On statistical grounds this procedure might be expected to lead to somewhat higher coefficients than if the data for the boys had been analyzed separately. On the other hand, in the Block, Block, and Harrington (1974) and Kohn and Rosman (1972) studies, small and negligible differences were found when the sta-

bility correlations were obtained separately for boys and girls and for the sexes combined. The children in these studies were also 3–4-year-olds. The results of the latter two studies thus suggest that combining the sexes in calculating a stability correlation for children of these ages may not have great effect on the size of the correlation coefficient.

Furthermore, there was consanguinity within the sample in one study (Kagan & Moss, 1962), and this fact may have to some extent inflated the raw stability correlations. However, as described in the unabridged report (Olweus, Note 1), certain countermeasures were taken in the present analyses to reduce or eliminate the risks of obtaining too high attenuation-correlated coefficients.

In some studies, partly overlapping, partly different sets of judges/raters were used for the different periods of time (Emmerich, 1966; Eron et al., 1972; Olweus, 1977a; Wiggins & Winder, 1961). However, the evidence presented by Olweus (1977a and p. 860 in the present article) suggests that the inflating effects may be negligible, at least when the time interval between the rating occasions is a year or more.

Potential Correlation-Lowering Factors

Several of the samples studied were probably relatively homogeneous as regards aggressive reactions and behaviors. Furthermore, in longitudinal studies there is usually sample shrinkage over time, and this shrinkage is often systematic. In particular, relatively aggressive individuals are most likely to disappear from the sample studied, as was found in the Eron et al. (1972) study. Such factors tend to reduce the size of the stability coefficients.

In the majority of studies, the conventional product-moment correlation coefficient was employed to express the degree of stability over time. A few studies (Jersild & Markey, 1935; Martin, 1964; Patterson et al., 1967), however, used Spearman's rho, and this index is likely to be slightly lower than the product-moment coefficient (Guilford, 1956). Furthermore, use of a somewhat different technique in forming total scores out of individual items very likely would have given higher stability correlations in at least one study (Wiggins & Winder, 1961).

A central problem in many longitudinal studies is that the conceptual equivalence of the variables or measurement instruments used at different periods of assessment may be considerably less than perfect. The implication is that if more equivalent variables or instruments had been employed, the stability correlations would have been higher. This problem applies to some of the studies considered in this article, for instance, to those in which different variables or instruments had to be used at different periods of time (Kagan & Moss, 1962; Kohn & Rosman, 1973). Generally, this problem seems to be particularly salient when the interval separating the times of assessment is long or the periods assessed represent very different developmental stages.

Finally, and perhaps most importantly, in several studies involving different school and nursery school classes (Block et al., 1974; Emmerich, 1966; Eron et al., 1972; Farrington, 1978; Kohn & Rosman, 1972, 1973; Wiggins & Winder, 1961), the stability correlations were calculated on the basis of the total samples and not as a (weighted) average of the within-class correlations. It can be shown that if the correlations between the means of the classes at the two times of measurement are lower than the (weighted) average of the within-class correlation, the result is a total correlation that is less than the average within-class correlation (cf. Lindquist, 1940, although no proof is given). Often there are good reasons to expect the correlation between the class means to be lower than the average within-class correlation (different raters or sets of raters may develop different, class-relative rating norms and so on), and accordingly, the total correlation will be an underestimate (for theoretical as well as predictive purposes the real interest is in a stability correlation unaffected by differences in class means). There are few reasons for expecting the correlation between the class means to be higher than the average within-class correlation in studies

of the present type. In the second (as well as the first) study by Olweus (1977a), results supporting the above argument were found. The raw correlations for the total sample ($N = 201$) were .55 for start fights and .63 for verbal protest, as compared with .65 and .70, respectively, when weighted average within-class correlations were computed (equivalent to total correlations based on within-class standardized variables). In other studies, the effects of eliminating class differences in mean level (and variability) may be less marked, but nevertheless, such a procedure is most likely to systematically increase the size of the stability coefficients.

All in all, the above considerations suggest that a preponderance of correlation-lowering factors were operative in the studies surveyed, and very likely, many of the stability coefficients reported are underestimates. In addition, when correcting for attenuation, a deliberate strategy was adopted to settle on reliability estimates that were not too low in cases of incomplete reliability information. Accordingly, the stability coefficients on which the following analyses are based are likely to be underestimates (rather than overestimates or "correct" values).

General Description of Results

If the disattenuated stability coefficients are plotted as a function of the interval in years between the two times of measurement ($T_2 - T_1$), a relatively regular picture is obtained (Figure 2). The size of the stability coefficient tends to decrease as the interval covered increases. The trend can be described by the following regression equation: $y = .78 - .018x$, where y is the disattenuated correlation and x is the interval ($T_2 - T_1$) in years. The spread around the regression line (the standard error of estimate, $s_{y \cdot x}$) is .13, and the correlation between the two variables amounts to −.66. The decrease of the regression line is relatively slow (although significant). For an interval of 5 years, the estimated disattenuated stability correlation is .69 and for an interval of 10 years is .60.[4]

For comparison, the linear regression line for data on intelligence test measurements (of the Stanford-Binet type) compiled by Thorndike (1933) is shown in Figure 2. (Thorndike's findings were later corroborated by other researchers and are widely accepted; see, e.g., Anastasi, 1958). In this case the regression line is based on 36 stability coefficients covering intervals up to 5 years. These coefficients were derived from 13 different samples, mainly school-age children (the ages were not given). The average sample size was 111. The regression equation describing the trend in Thorndike's data is $y = .92 - .022x$.[5] The standard error of estimate, $s_{y \cdot x}$, is .08, and the correlation between the variables amounts to −.32 (all measures were recomputed from Thorndike's data and corrected for attenuation; the assumed reliability was .95, in accordance with the first two entries in Thorndike's table). As evident from the figure and the regression line, the stability of intelligence measures is generally somewhat higher than the stability of aggression variables, but the decrease of the regression line is slightly steeper for intelligence than for aggression. Extrapolating from the regression line for intelligence, the difference between the estimated coefficients would be only .10 for an interval of 10 years (.70 for intelligence and .60 for aggression. Although it may not be feasible to institute very detailed comparisons between the two sets of data, since the ages of the subjects were not given in Thorndike's article, it can be generally asserted that there is a substantial degree of stability over time for aggression[6] as well as for intelligence and that the difference in stability

[4] In calculating the regression equation each stability coefficient was given equal weight. Weighting of the stability coefficients according to the number of subjects on which they are based makes little difference. The same is true for the use of Z values instead of r values. For the raw correlations (unweighted) the regression equation is $y = .63 - .014x$.

[5] The regression equation for the (unweighted) raw correlations is $y = .87 - .020x$.

[6] It can be noted in this context that the stability of aggressive reaction patterns in females also seems to be substantial (if two studies with relatively small samples are excepted), in contrast with what is generally assumed (manuscript in preparation).

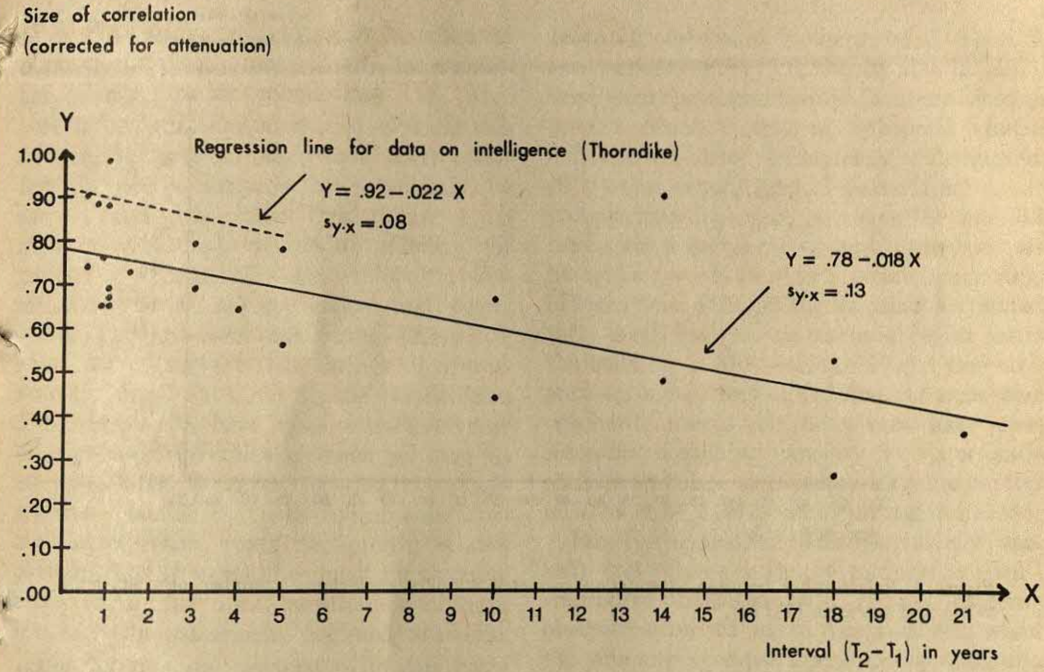

Figure 2. Regression line showing relationship between attenuation-corrected stability coefficients and time interval ($T_2 - T_1$) in years (unbroken line). (The regression line is based on 24 stability coefficients [plotted]. For comparison, the regression line for attenuation-corrected stability coefficients in the area of intelligence is shown [broken line]. This regression line is based on 36 stability coefficients [not plotted]; Thorndike = Thorndike, 1933.)

in the two areas does not appear to be very great.[7]

Although the goodness of fit of the regression line to the data in Figure 2 appears quite acceptable, theoretical considerations and inspection of Figure 1 suggest that the subjects' age at the time of first measurement may be an important parameter in addition to the interval $T_2 - T_1$. One way of expressing the stability correlations as a joint function of the subjects' age and the interval is to form an age ratio T_1/T_2. For a constant age, the age ratio decreases as the interval $T_2 - T_1$ increases. And for a particular interval, the age ratio is lower at a low age than at higher ages. It is theoretically reasonable to expect the stability coefficients to show the same trend as the age ratio, and accordingly, a positive relationship can be anticipated between the age ratio and the disattenuated stability correlations.

The relationship between the two variables shown in Figure 3 is clearly positive, as expected, and the goodness of fit of the regression line ($y = .26 + .617x$) to the data is even better than in Figure 2. This is manifested in a somewhat lower standard error of estimate (.11 compared with .13) or,

[7] In another overview by Thorndike (1940), the subjects in his 1933 article are referred to as school-age children. If in order to make the present data and the intelligence test data as comparable as possible, the regression equation is calculated only for aggression studies on children of school-age and for intervals up to 5 years (the first eight studies in the second section of Table 1), the equation is found to be quite similar to the regression equation for the total set of studies: $y = .80 - .027x$. It can also be noted that if comparisons are made on the preschool level, that is, comparisons concerning the stability of aggressive and intelligence test behavior (Thorndike, 1940, Table 1) over intervals up to a year or so for 3–4-year-old children, the stability of aggressive behavior is found to be as great as or slightly greater than the stability of intelligence test behavior (attenuation-corrected coefficients).

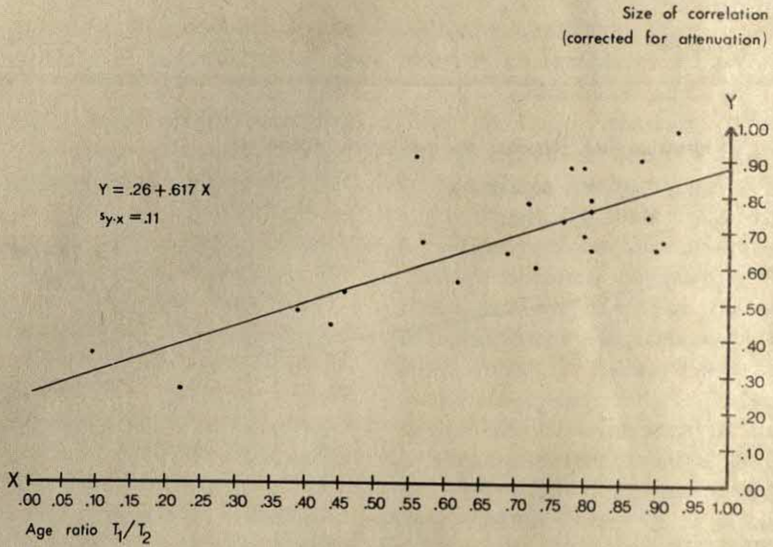

Figure 3. Relationship between attenuation-corrected stability coefficients (number of coefficients = 24) and age ratio (T_1/T_2).

alternatively, in a somewhat higher correlation coefficient (.77 compared with .66).[8] The picture obtained in Figure 3 indicates a good deal of regularity in the relationship between the age ratio T_1/T_2 and the disattenuated stability coefficients (similar results were obtained for the raw correlations). More specifically, the figure shows that the degree of stability in the individuals' relative positions on aggression variables of the type considered here can be broadly described as a linear positive function of the interval between the times of measurement and the subjects' age at the time of the first measurement, expressed in the ratio T_1/T_2.

There is thus less stability or more change in the individuals' relative positions the longer the interval $(T_2 - T_1)$ covered is (particularly if the subjects' age at T_1 is held constant). And for a particular interval $(T_2 - T_1)$, there is less stability or more change the younger the subjects are (although the relation between the stability correlation and the subjects' age is less marked than that between the stability correlation and the interval). The effect, so to speak, of a certain interval is greater for younger than for older subjects. These conclusions appear quite reasonable from a developmental as well as from a common sense point of view.

Since the ages of the subjects were not given in Thorndike's (1933) article, it is not possible to compare the fit of the regression lines for interval versus stability coefficients with the fit of the regression lines for age ratio versus stability coefficients. However, data available from other sources (Anastasi, 1958; Honzik, 1938) suggest that use of the age ratio T_1/T_2 in the intelligence domain, leads to better predictions than the use of the interval $T_2 - T_1$ alone.

More Specific, Descriptive Conclusions

The substantial degree of regularity manifested in Figure 2 and in Figure 3 is particularly impressive considering the great variation among the studies in sample composition, definition of variables, research setting, method of data collection and integration, and the researcher's theoretical orientation. There was also a very great range in the ages and intervals studied. After having emphasized the regularity of the data

[8] The correlation of .77 is in fact slightly higher than the multiple correlation between disattenuated coefficients on one hand and age and interval on the other; besides, use of the age ratio is more meaningful from a theoretical point of view.

as a general finding, it is appropriate to examine the results more closely for a number of more specific conclusions.

It is obvious that marked individual differences in habitual aggression level manifest themselves early in life (certainly by age 3) and may show (see Figure 1) a high or very high degree of stability for periods of at least 1½ years at this developmental level (in nursery school and school settings). Data from one study (Kagan & Moss, 1962) suggest that ratings of aggression variables that refer to the period from 0 to 3 years may have some predictive value of aggression variables assessed as long as 20 years later. However, to what extent aggressive reaction patterns observable during the preschool years can predict related patterns 5 or 10 years later must for the time being remain an open question, since data for such an assessment are not available.

Furthermore, in contrast with the common belief that the method of direct observation gives evidence of much more behavioral specificity and less stability than ratings of different kinds, no such tendencies were found in the present material. The average stability correlation for the three studies using direct observation (Jersild & Markey, 1935; Martin, 1964; Patterson et al., 1967) was .81, which can be compared with the average value of .79 for the three comparable studies by Block et al. (1974), Emmerich (1966), and Kohn and Rosman (1972; first study in Table 1) employing teacher ratings (the average stability correlations for the two sets of studies, using uncorrected coefficients, were .65 and .64, respectively). The comparability of these two sets of studies is manifested in equal average age ratios: both were .83; all studies were based on nursery school or similar groups. Judging from these studies, there seems to be no difference in degree of stability over relatively limited periods of time (up to a year) for aggression data collected by means of direct behavioral observation and teacher ratings.

Passing on to the school years, it is obvious that aggressive reaction patterns observable at ages 8 or 9 can be substantially correlated with similar patterns observed 10 to 14 years later (some 25% of the variance accounted for). It should also be noted that such patterns can with some success predict certain forms of antisocial violent behavior (violent delinquency; Farrington, 1978) that occur 10 to 12 years later.

Aggressive behavior at ages 12 and 13 may show a high or very high degree of stability for periods of 1 to 5 years (from 50% to more than 90% of the variance accounted for). Also, for periods as long as 10 years the stability is high (some 45% of the variance accounted for). Furthermore, aggressive reaction patterns at these ages have considerable predictive capacity for later antisocial aggression, as evidenced by the studies of Eron et al. (1972) and Farrington (1978).

Finally, aggressive behavior (chiefly verbal) and reactivity in the mid-30s are substantially correlated with similar patterns observed some 15 to 18 years earlier, when the subjects were teenagers. Considering the general trend of the stability coefficients and the fact that 52 of the 72 subjects used by Tuddenham (1959) were included in the sample of 171 subjects studied by Block (1971), it is likely that the disattenuated coefficient of .91 reported by Tuddenham is somewhat high because of chance factors. It should be noted, however, that the follow-up data in Tuddenham's study were completely independent of the data on which the adult evaluations in Block's study were based.

When evaluating these results, the general adequacy and validity of the data should also be considered. One should recall that in several investigations a considerable degree of correspondence was found between the aggression variables studied and teacher ratings of the same or related behaviors. This was true for teacher ratings and nominations versus peer ratings (Olweus 1974, 1978; Walder, Abelson, Eron, Banta, & Laulicht, 1961; Wiggins & Winder, 1961) as well as for teacher ratings versus direct behavioral observation (Jersild & Markey, 1935). If these sets of data were corrected for attenuation, the correlation between them very likely would exceed .75, indicating a quite substantial relationship. In some investigations the aggression variables studied also mani-

fested relationships of considerable magnitude with self-report data on similar patterns (Olweus, 1973b, 1978; Olweus, Note 2) and related, but more antisocial forms of behavior (Eron et al., 1972; Farrington, 1978). In addition, clear associations were obtained between two of the peer nomination instruments used in the stability studies and overt aggressive behavior in a contrived, naturalistic setting (Winder & Wiggins, 1964) and in a controlled, experimental situation, respectively (Williams, Meyerson, Eron, & Selmer, 1967). Finally, the possible existence of rater biases and stereotypes was carefully examined in some studies, in particular those by Block (1971) and Olweus (1977a). In the latter study it was concluded on the basis of several different analyses that "the rating data to an overwhelming degree reflect characteristics of the boys under study, rather than the biases and cognitive schemas of the raters irrespective of ratee characteristics" (Olweus, 1977a, p. 1310).

All in all, the above results derived by different methods and under a wide variety of conditions constitute strong evidence for the validity and general adequacy of the aggression data on which the stability correlations were based. They also attest to a substantial degree of cross-situational consistency in the sense that there is a considerable correspondence between aggression data obtained from independent sources or modes of measurement at about the same point in time. (The issue of cross-situational consistency in the area of aggression is not pursued further in the present article.)

It should also be noted that the finding of a considerable degree of stability of aggressive reaction patterns over time seems to be in general agreement with what has been observed in a number of studies of related, but more clearly antisocial forms of behavior (e.g., Conger & Miller, 1966; McCord & McCord, 1959; Robins, 1966; Roff, 1961; Rutter, 1972; Tait & Hodges, 1962).

Interpretation of the Stability Data

The descriptive conclusion that there is a substantial degree of stability in aggressive behavior cannot, however, without further analyses be taken as evidence for the corresponding stability of some reaction tendencies or motive systems within individuals. It might be argued from a situationist point of view (see Bowers, 1973), for instance, that the observed consistency primarily reflects stably different conditions for different individuals in the settings studied. Thus, in the first place, the stated conclusion can be said to apply under typical conditions, that is, under a degree of environmental variation (or stability) and pressure for change (or nonchange) typically found in the settings of the subjects for the periods studied (cf. Olweus, 1977a). Accordingly, it is important to examine the conditions characterizing the settings and periods under study, maybe particularly for the highly aggressive individuals, since their relative lack of change is a prerequisite to high stability coefficients.

In the studies on preschool children little detailed knowledge of changes in the settings is available. However, a good deal of change in the composition of the peer group took place in some studies (e.g., Jersild & Markey, 1935; Kohn & Rosman, 1972, 1973). Furthermore, in the three studies using the method of direct observation (see Table 1), the behavior was observed during periods of free play involving a minimum of situational structure and, in all probability, interaction with a number of different peers. Even if evidence has been presented that a nursery school setting can provide reinforcement of aggressive behavior (Patterson et al., 1967), the same authors have also reported (p. 32) that the highly aggressive children were the most likely to be the target of other children's aggression, that is, to get punished. Furthermore, Jersild and Markey (1935, p. 163) reported that the nursery school teachers interfered with the children's conflicts in about a third of the cases and predominantly in a way that was unfavorable to the aggressor. It should also be recalled that a small number of children accounted for a large percentage of the aggression episodes. It thus seems difficult to explain highly aggressive behavior in these settings as a consequence of situational pull

or primarily as a function of reinforcement. It appears rather that the highly aggressive children in particular were exposed to a certain pressure for change in nonaggressive directions from the teachers.

As regards the subject groups studied in the school setting it can be mentioned that in most cases the classes had new teachers and had moved to other school buildings at the second time of measurement. A relatively large percentage (10% to about 25%) of the original classmates had also been replaced by new peers. In one of the studies (Olweus, 1977a), somewhat more detailed information about changes in the composition of the peer group was available. As previously shown, even very marked changes in two classes did not affect the stability of aggressive behavior of those boys who were in the class at both times of measurement.

Furthermore, it can generally be assumed that there was at least a certain amount of pressure from the teachers and the administrative staff in the direction of modifying the behavior of the habitually aggressive pupils. As evidenced by the substantial stability correlations, such environmental pressure did not seem to be very effective. This finding is in good agreement with the general experience that it is difficult to reduce aggressive and antisocial behavior in preadolescent and adolescent males (see, e.g., Olweus, 1978, chap. 9; Burchard & Harig, 1976). It thus appears that the behavior of highly aggressive boys of these ages is often maintained irrespective of considerable environmental variation and in opposition to forces acting to change this same behavior.

It may be questioned, however, if there are not particular aversive situations or conditions in the school environment of the habitually aggressive boys that might explain their behavior. This question was analyzed in some detail by Olweus (1977a), who drew on the extensive findings regarding a particular group of highly aggressive boys, the bullies previously mentioned (Olweus, 1974, 1978). On the basis of several lines of evidence concerning the possible existence of frustrations, failures, and rejections in the school as well as the presence of other psychological, physical, and socioeconomic conditions of the bullies, it was concluded that "it is very difficult to explain the behavior of the highly aggressive boys as consequence of their being exposed to unusually aversive situations or conditions in the school setting" (Olweus, 1978, p. 136).

With regard to the subjects followed up in their adult years (Block, 1971; Kagan & Moss, 1962; Tuddenham, 1959), the majority went through experiences of potentially great impact on their lives during the interval from earlier to later measurements. Most of the subjects had married, and a certain percentage had divorced; in addition, they had started on and also covered part of their professional careers. A majority of the subjects in Block's study (and probably also in that of Tuddenham) had become parents. Most of them had been in military service, and roughly half of Block's (and the majority of Tuddenham's) subjects had served in the armed forces during World War II. Some of the situations or life events mentioned can be primarily regarded as forced upon the subjects, at least in some respects; others can be mainly considered to be a result of active selection on the part of the subjects. In sum, it is very likely that a good deal of environmental change and also pressure for change of highly aggressive reaction patterns were imposed on the subjects in these studies during the 10 to 20 years separating the earlier and later assessments. In addition, considerable maturational changes can be expected to occur during such a long period for subjects who are only about 10–15-years-old at the time of the early measurement.

When making an overall evaluation of actual and presumed environmental changes and pressures for change during the periods studied, one is, in fact, even more surprised at the degree of stability manifested. As previously concluded, changes in the individuals' relative positions had certainly occurred, both as a function of the interval covered and of the individuals' age at the time of first measurement. And if more detailed knowledge of the conditions and life events facing the individual subjects had been available, maybe more exact predictions about changers and nonchangers could have

been made. In an overall appraisal, however, the primary task confronting the researcher seems to be one of explaining the substantial stability or lack of change in aggressive behavior found to prevail in spite of considerable environmental variation and in opposition to a number of influences acting to change this same behavior.

The relative lack of change is all the more remarkable because highly aggressive behavior often leads to aversive consequences from the environment. Even if psychological and physical advantages can be gained by aggressive behavior in a number of situations, negative effects (such as punishment from the environment) often seem to be equally likely. It can also be argued that many aggressive behaviors, such as bullying or aggressive attacks in a free-play situation, are self-initiated behaviors (cf. Olweus, 1977a). As previously pointed out, it is often difficult to explain the behavior of the highly aggressive individuals as a function of particular aversive conditions or strong situational pull in the immediate, proximal situation in which the aggressive behavior is displayed. In the studies surveyed, there is little evidence supporting a view that stable differences in aggression level are primarily a consequence of consistently different environmental conditions for different individuals in the nursery school, the elementary school and so on. Overall, the above results and analyses strongly suggest that the observed stability over time of aggressive reaction patterns is, to a considerable measure, determined by relatively stable, individual-differentiating reaction tendencies or motive systems within individuals.

Conclusions

In addition to the previous descriptive generalizations, the following conclusions are warranted. They pertain directly to the issues raised in the Introduction.

1. The degree of consistency over time in aggressive behavior is much greater than has been maintained by proponents of a behavioral (situational) specificity position in the personality field (e.g., Mischel, 1968, 1969). It should be noted that the aggressive behavior and reaction patterns studied were observed or inferred by individuals other than the subjects themselves and that several studies (Jersild & Markey, 1935; Martin, 1964; Patterson et al., 1967) used the method of direct behavioral observation. It should also be emphasized that, generally, the substantial degree of stability found can hardly be interpreted as mainly reflecting consistency constructed in the minds of the observers, irrespective of the actual behavior of the subjects.

The across-time stability of aggressive behavior was not much lower than that typically found in the intelligence domain. This finding is worthy of particular emphasis, since the stability of behaviors associated with intelligence and cognitive processes has been generally regarded as impressive and indicative of "genuine continuity" also by proponents of a behavioral specificity position (e.g., Mischel, 1968, pp. 35–36). To avoid misunderstanding, however, I want to make clear that when pointing to similarities between results from the intelligence domain and those from the aggression area, I restrict my comparison to the degree of stability over time. I am in no way implying assumptions about similar developmental and operating mechanisms or, for instance, that the degree of genetic influence is the same in the two areas (see Olweus, 1978, chap. 8).

2. As previously spelled out, the results and analyses strongly suggest that important determinants of the observed consistency in aggressive behavior over time are to be found in relatively stable, individual-differentiating reaction tendencies or motive systems, however conceptualized, within individuals. This conclusion should not be taken to imply that situational factors are considered unimportant for the evocation of aggressive behavior (see, e.g., Olweus, 1969, 1973b). Nor does it imply that aggressive behavior is independent of rewarding and maintaining conditions in the immediate, proximal environment. However, it is contended here that the explanatory and predictive value of such factors has been exaggerated in the last decade; the analyses of the present article clearly suggest that relatively stable, internal reaction tendencies are important determinants

of behavior in the aggressive-motive area and should be given considerably greater weight than has been done recently. (For an overview of how such internal reaction tendencies may develop in the area of aggression, see Olweus, 1978, chap. 8.) In line with this argument, it also seems quite reasonable to assume that the inferred, internal reaction tendencies or motive systems within an individual are essential codeterminants of what the individual will perceive as reinforcing. In fact, the analyses presented suggest that highly aggressive individuals to a considerable degree actively select and create the kind of situations in which they are often observed (cf. Bowers, 1973; Wachtel, 1973).

The stated conclusion can also be interpreted as providing support for some form of trait position,[9] at least in the following sense: The results indicate that the probability of giving an aggressive response in potentially aggression-activating situations, more or less separated in time, or the strength of such responses differs greatly among individuals. In my view, however, the results do not imply that knowledge of an individual's habitual aggression level necessarily leads to good predictions of the behavior of the individual in a particular concrete situation. It has been empirically found, for instance, that the relationship between aggressive responses in different situations (data sources) for certain groups of individuals (with high aggression-inhibitory tendencies) may even be negative (but in a predictable way; see Olweus, 1969, 1973b). To make more accurate predictions for particular situations, it seems necessary to take into account, among other things, the individual's cognitive appraisal of the situation, the aggressive activation value of the situation, the aggression-inhibitory activation value of the situation, the strength of the individual's habitual aggressive tendencies, as well as the strength of the individual's aggression-inhibitory tendencies (Olweus, 1969). Accordingly, I prefer not to interpret the consistency results obtained in terms of a (simple) trait formulation of aggressiveness (see also, Olweus, 1973b).

The preceding analyses and conclusions thus indicate that what are known as personality concepts involving relatively stable, internal reaction tendencies or properties of individuals are useful in predicting and explaining aggressive behavior. Data on longitudinal consistency in other motive systems (e.g., Block, 1971) suggest that this is true also in areas of psychology other than aggression. It appears, then, in contrast with some recent proposals (e.g., Krasner & Ullmann, 1973; Shweder, 1975), that personality concepts and variables referring to relatively stable, individual-differentiating reaction tendencies or properties may be of great value in psychology for many years to come.

[9] It should be noted that it is difficult to speak of trait theory in general, without reference to a particular theorist or a particular motive area. There are obviously many differences and nuances among different theorists.

Reference Notes

1. Olweus, D. *Longitudinal studies of aggressive reaction patterns in males: A review* (Report No. 2). Bergen, Norway: University of Bergen, Institute of Psychology, 1977.
2. Olweus, D. *Development of a multi-faceted aggression inventory for boys* (Report No. 6). Bergen, Norway: University of Bergen, Institute of Psychology, 1975.

References

Anastasi, A. *Differential psychology*. New York: Macmillan, 1958.

Block, J. *The Q-sort method in personality assessment and psychiatric research*. Springfield, Ill.: Charles C Thomas, 1961.

Block, J. The equivalence of measures and the correction for attenuation. *Psychological Bulletin*, 1963, *60*, 152–156.

Block, J. *Lives through time*. Berkeley, Calif.: Bancroft Books, 1971.

Block, J. Advancing the psychology of personality: Paradigmatic shift or improving the quality of research? In D. Magnusson & N. S. Endler (Eds.), *Personality at the cross-roads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.

Block, J., Block, J. H., & Harrington, D. M. Some misgivings about the Matching Familiar Figures Test as a measure of reflection–impulsivity. *Developmental Psychology*, 1974, *10*, 611–632.

Bowers, K. S. Situationism in psychology: An analysis and a critique. *Psychological Review*, 1973, *80*, 307–336.

Burchard, J. D., & Harig, P. T. Behavior modification and juvenile delinquency. In H. Leitenberg (Ed.), *Handbook of behavior modification.* Englewood Cliffs, N. J.: Prentice-Hall, 1976.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81–105.

Conger, J. J., & Miller, W. C. *Personality, social class, and delinquency.* New York: Wiley, 1966.

Emmerich, W. Continuity and stability in early social development: II. Teacher ratings. *Child Development,* 1966, *37,* 17–27.

Endler, N. S., & Magnusson, D. (Eds.). *Interactional psychology and personality.* Washington, D. C.: Hemisphere, 1976.

Epstein, S. Traits are alive and well. In D. Magnusson & N. S. Endler (Eds.), *Personality at the cross-roads: Current issues in interactional psychology.* Hillsdale, N.J.: Erlbaum, 1977.

Eron, L. D., Huesmann, L. R., Lefkowitz, M. M., & Walder, L. O. Does television cause aggression? *American Psychologist,* 1972, *27,* 253–263.

Farrington, D. P. The family backgrounds of aggressive youths. In L. Hersov, M. Berger, & D. Schaffer (Eds.), *Aggression and antisocial disorder in children.* Oxford, England: Pergamon Press, 1978.

Golding, S. L. Flies in the ointment: Methodological problems in the analysis of the percentage of variance due to persons and situations. *Psychological Bulletin,* 1975, *82,* 278–288.

Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1956.

Honzik, M. P. The constancy of mental test performance during the preschool period. *Journal of Genetic Psychology,* 1938, *52,* 285–302.

Jersild, A. T., & Markey, F. V. Conflicts between preschool children. *Child Development Monograph,* No. 21, 1935.

Jones, H. E. The California adolescent growth study. *Journal of Education Research,* 1938, *31,* 561–567.

Kagan, J., & Moss, H. A. *Birth to maturity: A study in psychological development.* New York: Wiley, 1962.

Kohn, M., & Rosman, B. L. A social competence scale and symptom checklist for the preschool child: Factor dimensions, their cross-instrument generality, and longitudinal persistence. *Developmental Psychology,* 1972, *6,* 430–444.

Kohn, M., & Rosman, B. L. Cross-situational and longitudinal stability of social-emotional functioning in young children. *Child Development,* 1973, *44,* 721–727.

Krasner, L., & Ullmann, L. P. *Behavior influence and personality.* New York: Holt, Rinehart & Winston, 1973.

Lindquist, E. F. *Statistical analysis in educational research.* New York: Houghton Mifflin, 1940.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Macfarlane, J. Studies in child guidance: I. Methodology of data collection and organization. *Monographs of the Society for Research in Child Development,* 1938, *3*(6, Whole No. 19).

Martin, W. E. Singularity and stability of the profiles of social behavior. In C. B. Stendler (Ed.), *Readings in child behavior and development.* New York: Harcourt, Brace & World, 1964.

McCord, W., & McCord, J. *Origins of crime.* New York: Columbia University Press, 1959.

Mischel, W. *Personality and assessment.* New York: Wiley, 1968.

Mischel, W. Continuity and change in personality. *American Psychologist,* 1969, *24,* 1012–1018.

Olweus, D. *Prediction of aggression: On the basis of a projective test.* Stockholm, Sweden: Skandinaviska Testförlaget, 1969.

Olweus, D. *Hackkycklingar och översittare: Forskning om skolmobbing.* Stockholm, Sweden: Almqvist & Wiksell, 1973. (a)

Olweus, D. Personality and aggression. In J. K. Cole & D. D. Jensen (Eds.), *Nebraska Symposium on Motivation, 1972* (Vol. 20). Lincoln: University of Nebraska Press, 1973. (b)

Olweus, D. Personality factors and aggression: With special reference to violence within the peer group. In J. de Wit & W. W. Hartup (Eds.), *Determinants and origins of aggressive behavior.* The Hague, The Netherlands: Mouton, 1974.

Olweus, D. Aggression and peer acceptance in adolescent boys: Two short-term longitudinal studies of ratings. *Child Development,* 1977, *48,* 1301–1313. (a)

Olweus, D. A critical analysis of the "modern" interactionist position. In D. Magnusson & N. S. Endler (Eds.), *Personality at the cross-roads: Current issues in interactional psychology.* Hillsdale, N.J.: Erlbaum, 1977. (b)

Olweus, D. *Aggression in the schools: Bullies and whipping boys.* Washington, D.C.: Hemisphere, 1978.

Patterson, G. R., Littman, R. A., & Bricker, W. Assertive behavior in children: A step toward a theory of aggression. *Monographs of the Society for Research in Child Development,* 1967, *32*(5, Whole No. 113).

Robins, L. N. *Deviant children grown up.* Baltimore, Md.: Williams & Wilkins, 1966.

Roff, M. Childhood social interactions and young adult bad conduct. *Journal of Abnormal and Social psychology,* 1961, *63,* 333–337.

Rutter, M. Relationships between child and adult psychiatric disorders: Some research considerations. *Acta Psychiatrica Scandinavica,* 1972, *48,* 3–21.

Shweder, R. A. How relevant is an individual difference theory of personality? *Journal of Personality,* 1975, *43,* 455–484.

Spearman, C. Correlations of sums and differences. *British Journal of Psychology,* 1913, *5,* 417–426.

Tait, C. D., & Hodges, E. F. *Delinquents, their families, and the community.* Springfield, Ill.: Charles C Thomas, 1962.

Thorndike, R. L. The effect of interval between test and retest on the constancy of the IQ. *Journal of Educational Psychology*, 1933, *24*, 543–549.

Thorndike, R. L. "Constancy" of the IQ. *Psychological Bulletin*, 1940, *37*, 167–187.

Thouless, R. H. The effect of errors of measurement on correlation coefficients. *British Journal of Psychology*, 1939, *29*, 383–403.

Tuddenham, R. D. The constancy of personality ratings over two decades. *Genetic Psychology Monographs*, 1959, *60*, 3–29.

Wachtel, P. L. Psychodynamics, behavior therapy, and the implacable experimenter: An inquiry into the consistency of personality. *Journal of Abnormal Psychology*, 1973, *83*, 324–334.

Walder, L. O., Abelson, R. P., Eron, L. D., Banta, T. J., & Laulicht, J. H. Development of a peer-rating measure of aggression. *Psychological Reports*, 1961, *9*, 497–556. (Monograph)

Wiggins, J. S., & Winder, C. L. The Peer Nomination Inventory: An empirically derived sociometric measure of adjustment in preadolescent boys. *Psychological Reports*, 1961, *9*, 643–677.

Williams, J. F., Meyerson, L. J., Eron, L. D., & Selmer, I. J. Peer rated aggression and aggressive responses elicited in an experimental situation. *Child Development*, 1967, *38*, 181–190.

Winder, C. L., & Wiggins, J. S. Social reputation and social behavior: A further validation of the Peer Nomination Inventory. *Journal of Abnormal and Social Psychology*, 1964, *68*, 681–684.

Received April 3, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 876–879

Evolutionary Scales Lack Utility: A Reply to Yarczower and Hazlett

John P. Capitanio and Daniel W. Leger
University of California, Davis

Yarczower and Hazlett have proposed that evolutionary scaling based on anagenesis (biological improvement) is an acceptable—even desirable—facet of contemporary comparative psychology. We strongly disagree with that thesis. Our criticisms are based on (a) their misconception of anagenesis, (b) inconsistencies in the use of the term *evolutionary grades,* (c) typological thinking, and (d) the lack of utility of evolutionary scales. Reversion to evolutionary scaling by comparative psychologists would disrupt the ongoing synthesis of comparative psychology with other evolutionary sciences.

In a recent *Psychological Bulletin* article, Yarczower and Hazlett (1977) attempted to legitimize the construction of evolutionary scales, suggesting that "evolutionary scales have a place within comparative psychology and do not violate principles of evolution when done properly" (p. 1096). Although evolutionary scaling may have a limited place in analyses of structural features, we believe the concept is inappropriate and potentially misleading in analyses of behavior. Further, it is our contention that the arguments and evidence promulgated by Yarczower and Hazlett are not only internally inconsistent but reflect a serious lack of understanding of the principles of evolutionary biology.

Our criticisms of Yarczower and Hazlett revolve about four main issues: (a) the concept of anagenesis, or biological improvement, (b) the appropriate definition of evolutionary grade, (c) typological thinking, and (d) the utility of evolutionary scales in modern comparative psychology.

Anagenesis

Although most evolutionary biologists agree in principle that anagenesis refers to the evolution of increased complexity in some trait, the term has become diluted since its inception and "has come to be applied to nearly any kind of evolutionary change, whether leading to a marked advance or not" (Dobzhansky, Ayala, Stebbins, & Valentine, 1977, p. 236). Dobzhansky et al. feel that the vast majority of evolutionary events are actually adaptive changes at the same level of complexity.

Yarczower and Hazlett apparently view increased complexity and specialization as corollaries of evolution; they cited Darwin and some older works of Simpson and of Huxley to support this position. Although the former authors have the right to restrict their discussion to anagenesis, it should be recognized that anagenesis is but one of several evolutionary processes. For example, Dobzhansky et al. (1977) have written that the definition of evolution

> must not include the notion that evolution is always progressive, leading inevitably from simpler to more complex forms of life. To be sure, some of the most important evolutionary events have been increases in structural complexity, e.g., . . . the complex sense organs of vertebrates . . . and animal societies . . . . Furthermore, both older and modern theories [of evolution] recognize even degenerations of structure as evolution, provided they are products of adaptive alterations in population-environment interactions. (p. 8)

The evolution of blindness in cave-dwelling fish and the degeneration or structure and behavior in parasites are examples of the adaptiveness of *decreased* complexity. Wilson

Requests for reprints should be sent to John P. Capitanio or Daniel W. Leger, Department of Psychology, University of California, Davis, California 95616.

(1975) has provided a detailed account of the degeneration of the central nervous system and of behavior in parasitic ants compared with their ancestral free-living relatives.

Thus evolutionary change is by no means unidirectional. Because selection pressures often fluctuate, a certain amount of evolutionary backtracking is to be expected. Natural selection can result not only in anagenesis but also in degeneration, in stasigenesis (no change whatsoever), or, more likely, in change without alteration of complexity. It should be remembered that the things of value for organisms are survival and reproduction; evolutionary success is clearly not dependent on complexity per se.

Evolutionary Grades

The term *evolutionary grade*, as used by Yarczower and Hazlett, differs markedly from the definitions given in current major works (e.g., Mayr, 1970; Wilson, 1975). Although all the above authors have agreed that grade refers to the stage of a behavior, anatomical structure, or physiological process, Yarczower and Hazlett (1977) stipulated that to be included in the same grade, animals must be "related by parallel evolution" (p. 1091), that is, must be closely related. But according to Wilson (1975), "Phylogenetically remote lines can reach and pass through the same grades, in which case we speak of the species making up these lines as being convergent with respect to the trait" (pp. 25–26).

The concepts of parallel evolution and convergent evolution, with their implications of closely and distantly related species, respectively, have been a common source of confusion. Kaster and Berger (1977) correctly pointed out that parallelism and convergence can occur in taxa of *any* degree of phylogenetic affinity; that is, the process is of primary importance, and the degree of relatedness is of secondary concern.

But even if one refrains from judging the relative value of the two definitions, Yarczower and Hazlett were at least repeatedly inconsistent in using the term *grade* as they originally defined it. For example, they wrote of birds, mammals, and reptiles as having attained the grade Amniota. By no stretch of the imagination can these classes be viewed as being closely related; there is no doubt however that they have converged on a common strategy. Later, Yarczower and Hazlett diluted their original statements by following up Huxley's (1958) suggestion that the "breadth" of the grade be designated by the lowest common taxon rank of the least related species included in the comparison, for example, grade–order or grade–class (see Kaster & Berger, 1977, for a similar suggestion). Thus, Yarczower and Hazlett defined evolutionary grade differently from Wilson (1975), Mayr (1970), and Dobzhansky et al. (1977); but their use of the term is consistent with the latter authors' definition.

If, for the moment, one ignores Yarczower and Hazlett's (1977) examples and takes their definition of grade as what they actually meant, then they are faced with two major problems. First, their notion of parallelism must be clarified, because the possibility exists that even phylogenetically remote taxa may evolve in parallel. Second, if parallelism is taken to refer only to "closely" related species, they must then consider the distinct probability that they were dealing with homologous traits—a task they explicitly wished to circumvent (p. 1090).

But if one ignores their definition and views their examples as indicative of their position, it is quite clear that Yarczower and Hazlett were dealing with evolutionary convergence, that is, similar phenotypic response to similar selection pressures. Again, however, convergence was explicitly excluded from their thesis.

The failure of Yarczower and Hazlett to unambiguously state their position concerning evolutionary grades can only be taken as an indication that they have failed to establish "a third approach to the study of behavioral evolution" (p. 1088); their approach cannot be distinguished from the other two approaches, namely, the studies of homology and analogy.

Typology

In discussing the level of a particular grade, Yarczower and Hazlett (1977) asked us to keep in mind the representativeness of a particular animal with respect to its higher taxon. Is the cat representative of Carnivora,

the squirrel of Rodentia, and so on? One might well ask, is any species truly representative of a particular higher taxon? The question of representativeness is surely a relative one, and disagreement is certain to exist. Consider an example (from Tinbergen, 1958). If one wished to construct grades based on nesting behavior, one might very well choose the black-headed gull to represent the gull family (Laridae) and then place the gulls in the grade *colonial ground nesters*. Alternatively, one might choose the kittiwake as the representative and therefore place the gulls in the grade *colonial cliff nesters*. To do either would be an injustice, for one would lose the sense of variability the other exemplifies. One could eliminate the problem by renaming the grade *colonial nesters*, but by doing so one would lose more of the appreciation for diversity within the family.

The example may be an extreme one, but it does not depict a rare occurrence. Closely related species can have wide variations in any trait—Consider the greatly different social systems of the baboon species *Papio hamadryas* (one-male harem) and *Papio anubis* (multimale troop). And depending on one's definition of relatedness, one finds that within the primate superfamily Hominoidea, a wide variety of social organizations exist—from monogamous pairs to troop-living species (Eisenberg, Muckenhirn, & Rudran, 1972). Picking a representative would certainly be an impossible task. The only mechanism in the grade scheme for dealing with typologies of this sort is to make the grade ever more general. Far from being useful, grades then become arbitrary, broad, and meaningless.

Utility of Evolutionary Scales in Comparative Psychology

Another serious flaw in the Yarczower–Hazlett formulation is their almost complete inattention to behavior, even though their stated purpose was to examine behavioral evolution. Virtually all their examples come from anatomy and physiology—areas in which complexity can be measured relatively simply. But where behavior is concerned, complexity is an intractable concept. To continue a previous example, on the basis of sheer numbers, it might be concluded that a multimale troop's social structure is more complex than that of a one-male harem. Alternatively, one might argue that the loose aggregation of *P. hamadryas* baboons is actually more complex in that it is composed of highly discrete one-male units that manage to cooperate in foraging, traveling, and so on, with few overt interactions. Similarly, which is more complex, polygyny or polyandry? Is scent marking a territory more complex than vocally advertising that territory? In short, when considering behavior one is usually hard pressed to decide exactly how to rank order the candidates in terms of increasing complexity.

When faced with the difficult task of ranking several behaviors, one should be aware of the temptation to simply call the behavior found in the more recently evolved taxon the more complex. In our opinion, Yarczower and Hazlett may have fallen into this trap when they claimed that probability-maximizing, a strategy adopted by the rats in Bitterman's (1965) classic studies, is more complex and advanced than probability-matching, the strategy used by fish. We feel that a strong case can also be made for probability-matching being the more complex behavior. Maximization requires only a greater-than/less-than comparison; but in matching, the degree of difference between the two manipulanda is taken into account. Also, as Bitterman's data suggest, matchers adjust their response distributions very quickly following changes in reinforcement distribution, reflecting a sensitivity to environmental change exceeding that of the maximization strategy. The lesson of this example should be clear: One must not assume a priori that the more recently evolved taxon will necessarily exhibit the more complex form of the behavior.

Given that one assumes Yarczower and Hazlett's rationale for the development of grades and then constructs a series of grades based on some trait, one is left with the final question, What can be said about the results? Unfortunately, we feel the final answer is, Nothing. At best one arrives at the description of behavior in related animals (after arbitrarily defining *related* and toiling over choosing representative examples). At

worst, however, one does something more nefarious—One gives the impression that the grades represent a *scala naturae*. This impression arises from two points: first, that there is a rough correlation between complexity and evolutionary recency, which tends to clump "higher" animals in the more "improved" grades, and second, that the actual presentation of description in this form must have some underlying reason for existence other than description. In short, little is accomplished, except perhaps the generation of a lively debate over which grade is really the most improved. Indeed, once a scale has been developed, the only possible use that can be made of it (beyond the aforementioned descriptive function) is to make inferences that must employ adaptation of the particular organisms to their respective ecological niches.

We vigorously disagree with the statement (Yarczower & Hazlett, 1977, p. 1092) that the construction of a hierarchical progression of grades can have evolutionary significance beyond that of the study of adaptations. The source of our disagreement stems from the fact that characters exist as adaptations. To speak only of characters *in vacuo*, divorced of their adaptive significance, is meaningless in terms of evolution. Thus, to say that reptiles, although having "attained" the grade Amniota, have not attained the grade Homeothermy, "which reflects an improvement in the mechanism responsible for the regulation of body temperature" (Yarczower & Hazlett, 1977, p. 1096), is to say nothing of significance. The concept of not attaining this grade is meaningless, since selection pressures on reptiles are such that poikilothermy is generally favored.

In conclusion, we believe that evolutionary scales have no place in a modern comparative science of behavior. The idea of improvement smacks of anthropocentrism and is certainly arbitrary. These concepts are without value and may be seriously misleading. Scaling can only hinder interpretation of comparative psychologists' efforts by more traditional, evolution-oriented behaviorists at a time when these disciplines are reaching common grounds for interaction after having been polarized for so long at the extremes of the learned–instinct continuum.

References

Bitterman, M. E. Phyletic differences in learning. *American Psychologist*, 1965, *20*, 396–410.

Dobzhansky, T., Ayala, F. J., Stebbins, G. L., & Valentine, J. W. *Evolution*. San Francisco: Freeman, 1977.

Eisenberg, J. F., Muckenhirn, N. A., & Rudran, R. The relation between ecology and social structure in primates. *Science*, 1972, *176*, 863–874.

Huxley, J. S. Evolutionary processes and taxonomy with special reference to grades. *University of Uppsala Arsskrift*, 1958, pp. 21–39.

Kaster, J., & Berger, J. Convergent and parallel evolution: A model illustrating selection, phylogeny, and phenetic similarity. *BioSystems*, 1977, *9*, 195–200.

Mayr, E. *Populations, species, and evolution: An abridgment of animal species and evolution*. Cambridge, Mass.: Harvard University Press, 1970.

Tinbergen, N. *Curious naturalists*. Garden City, N.Y.: Doubleday, 1958.

Wilson, E. O. *Sociobiology: The new synthesis*. Cambridge, Mass.: Belknap Press, 1975.

Yarczower, M., & Hazlett, L. Evolutionary scales and anagenesis. *Psychological Bulletin*, 1977, *84*, 1088–1097.

Received April 10, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 880–883

In Defense of Anagenesis, Grades, and Evolutionary Scales

Matthew Yarczower and Bret S. Yarczower
Bryn Mawr College

We show that the first three criticisms by Capitanio and Leger of the Yarczower and Hazlett article are unfounded and that the fourth is premature: (a) Contemporary leaders in the study of evolution define anagenesis in the same way as did Yarczower and Hazlett; (b) their use of the term *evolutionary grade* was internally consistent and consistent with usage by Mayr, whom Capitanio and Leger cited as having used it differently; (c) classification of species into higher taxa does not represent "typological thinking"; and (d) although analyses by grades of social behaviors are more difficult than those of sensory systems, as was noted by Yarczower and Hazlett, it is premature to conclude that the effort will not bear fruit.

We answer in turn each of the four criticisms leveled at the Yarczower and Hazlett (1977) article and show that the first three are unfounded and that the fourth is premature.

1. *"Misconception of anagenesis."* Capitanio and Leger (1979) claimed that although at one time anagenesis meant what Yarczower and Hazlett claimed it to mean, it currently refers to any evolutionary change.

Gould (1976), a paleobiologist at Harvard's Museum of Comparative Zoology, has written about the concepts of anagenesis, evolutionary progress, and grades in a book that was received too late for the ideas contained within it to be incorporated into the Yarczower and Hazlett article. He wrote, "The standard evolution tree . . . leaves out evolutionary progress (or anagenesis) entirely . . . . Grades . . . are successive levels of organization defined as stages in the *improvement* of an organic design for some specified function" (p. 117; italics added). Or read what Jerison, author of *Evolution of the Brain and Intelligence* (1973), had to say about the terms *anagenesis* and *grades* in his address entitled "Smart Dinosaurs and Comparative Psychology" (Note 1). He said, "For an evolutionary analysis of intelligence, we might seek evidence of 'intellectual' *progress* from earlier to later species. This analysis is called 'anagenetic' and is about *progressive evolution*" (pp. 1–2; italics added). And again, "In anagenetic analysis the objective is to identify *grades* in evolution, recognizing the possibility of higher, or more advanced, grades on a particular dimension" (p. 2).

In fact, if one reads further in the source used by Capitanio and Leger to define anagenesis (Dobzhansky, Ayala, Stebbins, & Valentine, 1977) one reads, "Anagenetic episodes commonly create organisms with novel characters and abilities *beyond* those of their ancestors" (p. 236; italics added). And again, in describing the relationship among anagenesis, cladogenesis, and stasigenesis, anagenesis is defined as "evolutionary *advance* or change" (p. 236; italics added). Thus, the definition of anagenesis in the Yarczower and Hazlett article is one shared by leading contemporary students of evolution.

2. *"Inconsistencies in the use of the term* evolutionary grades." Capitanio and Leger's (1979) first criticism, although incorrect, at least is clear; the second is unclear. We believe there are two points they wished to make. They wrote that if "parallelism is

Requests for reprints should be sent to Matthew Yarczower, Department of Psychology, Bryn Mawr College, Bryn Mawr, Pennsylvania 19010.

taken to refer only to 'closely' related species, they must then consider the distinct probability that they were dealing with homologous traits—a task they explicitly wished to circumvent" (p. 877). What Yarczower and Hazlett (1977) wrote, in fact, was that "in the study of anagenesis it is important that the animals be related at least by parallel evolution" (p. 1090). It is puzzling that Capitanio and Leger read this and concluded that homologous traits were meant to be excluded from an anagenetic analysis. Animals must be related at least by parallel evolution, and obviously, direct descendancy or relationships underlying behavioral homologies satisfy the criterion of being related "at least by parallel evolution."

Capitanio and Leger also suggested that Yarczower and Hazlett's use of the term *evolutionary grade* differed from that of Mayr (1970) and Wilson (1975). Mayr (1963) wrote, "The felicitous term 'grade' was introduced into the evolutionary literature by Huxley . . . following Simpson . . . to designate 'a step of anagenetic advance, or unit of biological improvement'" (pp. 608–609). In a later abridged version of the same work, Mayr (1970) offered exactly the same sentence, but omitted Simpson's name. In any case, in both versions, Mayr then went on to note that "several *related* lines may reach the same adaptive or structural grade independently" (1963, p. 609; 1970, p. 365). Further, Mayr (1970) wrote that "Simpson in particular . . . has pointed out how rapidly a new type may reach a new phylogenetic 'grade,' but once this grade is reached the type remains essentially stable" (pp. 370–371; the 1963 version on p. 617 is almost identical). Mayr's discussion of grades relied heavily on Simpson's (1961) usage, as did Yarczower and Hazlett's (1977, pp. 1091–1092), and thus it should come as no surprise that usage of the term *evolutionary grade* by Mayr (1963, 1970) and by Yarczower and Hazlett does not differ and that no evidence to the contrary was presented by Capitanio and Leger. Wilson's (1975) suggestion "that phylogenetically remote lines can reach and pass through the same grades" (p. 25) does differ in principle from the suggestion of Yarczower and Hazlett (1977), as well as from that of Simpson (1961) and that of Mayr (1970).

3. *"Typological thinking."* Capitanio and Leger applied the concept of typology, or typological thinking, inappropriately. For its correct use, see Mayr (1963, 1969, 1970, 1976). We answer, however, the criticism raised by them. They seemed to object to the placing of two instances into a single class. They claimed that the uniqueness and individuality of each instance is lost when one notes the similarity between the two instances. Yet, surely statements about commonalities among instances are a goal of science. Indeed, when Capitanio and Leger labeled the gulls in their example as *kittiwakes,* they could very well have been accused of having lost an appreciation of the differences between Judy and Fred Kittiwake. Yarczower and Hazlett noted that inclusion in one grade did not mean that the same groups of animals would be placed together in other grades. The diversity among groups of animals indeed was recognized. But Capitanio and Leger appear to object to classification of species into higher taxa. If they do, then surely they must object to an important goal of the field of systematics. Mayr (1969) wrote that "each species may exist in numerous forms (sexes, ages, classes, seasonal forms, morphs, and other phena). It would be impossible to deal with this enormous diversity if it were not ordered and classified" (p. 1). And again, "One of the major preoccupations of systematics is to determine . . . what unique properties of every species and *higher taxon* are. Another is to determine what properties certain taxa have in common with each other" (p. 3; italics added).

4. *"Lack of utility of evolutionary scales."* This final charge reflects a number of misunderstandings by Capitanio and Leger as well as a legitimate challenge, one that Yarczower and Hazlett issued themselves. Capitanio and Leger (1979), in discussing social behavior, suggested that "one is usually hard pressed to decide exactly how to rank order the candidates in terms of increasing complexity" (p. 878). Yarczower and Hazlett (1977) stated explicitly that "it is more difficult to obtain agreement about what con-

stitutes improvement in social systems than about sensory systems. However, difficulty alone is not sufficient grounds for rejecting the notion of evolutionary scales" (p. 1096). Incidentally, Capitanio and Leger appear to treat increased complexity as synonymous with progressive improvement, but it should be noted that in their discussion of an example of progressive improvement in color vision, Yarczower and Hazlett never used the word *complexity*. This word was used in a brief review of the history of the concept of anagenesis. It is difficult to define improvement but not impossible. Consider an analysis of facial behavior. If facial movement is treated as the behavioral character to be subjected to an analysis by grades, then it is not unreasonable to assume that improvement is reflected in the increased ability to engage in greater varieties and intensities of facial behaviors. It is clear that the evolution of facial musculature is of prime importance in understanding the progressive improvement that seems to be reflected in the evolution of facial behavior (Chevalier-Skolnikoff, 1973; Huber, 1931/1972). The social significance of the evolution of facial musculature and the concomitant improvement in facial movements and facial behavior are considerable. The importance of research on facial behavior and facial expression (e.g., Ekman, 1973; Ekman & Friesen, 1976; Izard, 1971) for an understanding of a rich variety of social behavior provides some optimism that it may be possible to define improvement in systems relevant to social phenomena.

Capitanio and Leger (1979) concluded that Yarczower and Hazlett discussed "characters *in vacuo*, divorced of their adaptive significance" (p. 879). They came to this conclusion from a statement by Yarczower and Hazlett (1977) that "a hierarchical progression of grades . . . has evolution significance beyond that of the study of adaptations" (p. 1092). Even a casual reading of this statement does not lead to the conclusion that the adaptive significance of the grades is to be ignored. Yarczower and Hazlett discussed the differences among the goals of the study of behavioral homologies, adaptations, and grades, and they need not be repeated here.

A legitimate question raised by Capitanio and Leger asks whether analyses by grades can provide any interesting or important answers. They concluded that they cannot. This question cannot be treated in isolation from other attempts to understand the evolution of behavior. For example, there is a lively controversy about whether the search for behavioral homologies is likely to be a profitable one (e.g., Atz, 1970; Hailman, 1976; Hodos, 1976). Will the analyses of grades be profitable? Mayr (1976) stated, "The existence of minor and major grades is one of the most interesting phylogenetic phenomena, even though it is a phenomenon which we are still unable to understand adequately" (p. 450). And again, "To the evolutionary taxonomist the existence of grades seems often more significant and more meaningful biologically than the mere splitting of phyletic lines" (p. 451). Gould (1976) noted that

> lemur-monkey-ape-man is a caricature of primate phylogeny; once we are sure that the species we study are representative of their grade, this same sequence may well unravel the mysteries of neocortical function in successive levels of organization of the primate brain. . . . I do not think that analysis by grades is likely to be abandoned even in a pipedream world where phylogenies are laid out upon laboratory tables. (p. 121)

It may well turn out that analysis by grades will not bear fruit, but better that this be the result of recognizing, testing, and rejecting the potential value of analysis by grades than the result of ignorance about grades' existence.

Reference Note

1. Jerison, H. J. *Smart dinosaurs and comparative psychology*. Paper presented at the meeting of the American Psychological Association, Toronto, Canada, August 1978.

References

Atz, J. W. The application of the idea of homology to behavior. In L. R. Aronson et al. (Eds.), *Development and evolution of behavior*. San Francisco: Freeman, 1970.

Capitanio, J. P., & Leger, D. W. Evolutionary scales

lack utility: A reply to Yarczower and Hazlett. *Psychological Bulletin*, 1979, *86*, 876–879.

Chevalier-Skolnikoff, S. Facial expression of emotion in nonhuman primates. In P. Ekman (Ed.), *Darwin and facial expression*. New York: Academic Press, 1973.

Dobzhansky, T., Ayala, F. J., Stebbins, G. L., & Valentine, J. W. *Evolution*. San Francisco: Freeman, 1977.

Ekman, P. *Darwin and facial expression*. New York: Academic Press, 1973.

Ekman, P., & Friesen, W. V. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1976, *1*, 56–75.

Gould, S. J. Grades and clades revisited. In R. B. Masterton, W. Hodos, & H. Jerison (Eds.), *Evolution, brain, and behavior: Persistent problems*. Hillsdale, N.J.: Erlbaum, 1976.

Hailman, J. P. Uses of the comparative study of behavior. In R. B. Masterton, W. Hodos, & H. Jerison (Eds.), *Evolution, brain, and behavior: Persistent problems*. Hillsdale, N.J.: Erlbaum, 1976.

Hodos, W. The concept of homology and the evolution of behavior. In R. B. Masterton, W. Hodos, & H. Jerison (Eds.), *Evolution, brain, and behavior: Persistent problems*. Hillsdale, N.J.: Erlbaum, 1976.

Huber, E. *Evolution of facial musculature and facial expression*. New York: Arno Press, 1972. (Originally published, 1931.)

Izard, C. E. *The face of emotion*. New York: Appleton-Century-Crofts, 1971.

Jerison, H. J. *Evolution of the brain and intelligence*. New York: Academic Press, 1973.

Mayr, E. *Animal species and evolution*. Cambridge, Mass.: Harvard University Press, 1963.

Mayr, E. *Principles of systematic zoology*. New York: McGraw-Hill, 1969.

Mayr, E. *Populations, species, and evolution: An abridgment of Animal species and evolution*. Cambridge, Mass.: Harvard University Press, 1970.

Mayr, E. *Evolution and the diversity of life*. Cambridge, Mass.: Harvard University Press, 1976.

Simpson, G. G. *Principles of animal taxonomy*. New York: Columbia University Press, 1961.

Wilson, E. O. *Sociobiology: The new synthesis*. Cambridge, Mass.: Harvard University Press, 1975.

Yarczower, M., & Hazlett, L. Evolutionary scales and anagenesis. *Psychological Bulletin*, 1977, *84*, 1088–1097.

Received December 1, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 4, 884–888

Protecting the Overall Rate of Type I Errors for Pairwise Comparisons With an Omnibus Test Statistic

H. J. Keselman
University of Manitoba
Winnipeg, Canada

Paul A. Games
Pennsylvania State University

Joanne C. Rogan
University of Manitoba, Winnipeg, Canada

Two procedures for protecting the number of false rejections for a set of all possible pairwise comparisons were compared. The two-stage strategy of computing pairwise comparisons, conditional on a significant omnibus test, was compared with the multiple comparison strategy that sets a "familywise" critical value directly. The analysis of variance test, the Brown and Forsythe test, and the Welch omnibus test, as well as three procedures for assessing the significance of pairwise comparisons, were combined into nine two-stage testing strategies. The data from this study establish that the common strategy of following a significant analysis of variance F with Student's t tests on pairs of means results in a substantially inflated rate of Type I error when variances are heterogeneous. Type I error control, however, can be obtained with other two-stage procedures, and the authors tentatively consider the Welch F''–Welch t'' combination desirable. In addition, the two techniques for controlling Type I error do not substantially differ as much as might be expected; some two-stage procedures are comparable to simultaneous techniques.

Given K independent samples of size n_k, many experimenters are interested in testing the equality of the pairs of means. One can consider the complete set of means a family, in that if H_0, $\mu_1 = \mu_2 = \mu_k = , \ldots, = \mu_K$, is true, then it is also true that $\mu_k = \mu_{k'}$ for all possible pairs. The risk of making one or more Type I errors on the pairs is identified as the "familywise" risk of Type I error (FWI). This is contrasted with the risk of making a Type I error on a single contrast, which is labeled the per-comparison Type I error rate (PCI). Two general procedures can be used in this situation. A simultaneous multiple comparison procedure such as Tukey's (Note 1) wholly significant difference test compares all pairs using a critical value (CV) that controls FWI.

This research was supported in part by Canada Council Grant 451-77065 and in part by the Pennsylvania State University Computation Center.

Requests for reprints should be sent to H. J. Keselman, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2.

The second, more likely used procedure is a two-stage strategy. The control of FWI is accomplished by an initial omnibus test such as the analysis of variance (ANOVA) F test. Only if this first stage is significant does the user proceed to test the pairs. Then the pairs can be tested by using a critical value that controls only PCI. Such a critical value is always smaller than the above FWI CV. Consequently, it automatically follows that if one reaches the second stage, the tests on pairs will have greater power than the tests on pairs that directly control FWI via a larger CV. However, since this second stage is equivalent to doing $K(K-1)/2$ pairs of t tests, the total risk of Type I error rises with K, so that only the first stage provides FWI control in this process. This procedure, commonly referred to as the protected least significant difference (LSD) technique, was introduced by Fisher (1949) and was recommended by Carmer and Swanson (1973) over most multiple comparison procedures.

Unfortunately, Carmer and Swanson used

only the omnibus ANOVA and investigated only the equal sample size and homogeneous variance condition in their study. The ANOVA is known to be sensitive when unequal sample sizes are combined with unequal population variances (Box, 1954), so that this omnibus test often does not provide acceptable control of FWI when sample sizes are not equal. However, two alternate omnibus tests for mean equality have been shown to be more robust than the ANOVA F (Brown & Forsythe, 1974; Kohr & Games, 1974).

The Brown and Forsythe test, like the ANOVA, uses just sample sizes in the numerator to weight means, whereas the multisample Welch (1951) test weights the means by sample variances as well as by sample sizes. However, both tests, rather than obtaining a CV based on the usual ANOVA error degrees of freedom (df), obtain a modified CV that is a function of sample variances and sizes. Consequently, either of these tests may provide the stable FWI control needed to make the two-stage procedure robust. However, Type I error control still may not be achieved when Student's t tests are used to assess the pairwise comparisons following a significant first-stage omnibus test, particularly when the omnibus test is the ANOVA; that is, prior literature (e.g., Box, 1954; Boneau, 1960) suggests that two-stage procedures using a pooled within-cell estimate of error variation in either stage should not maintain the Type I error rate at the significance level when variances are heterogeneous.

Type I error control may be achieved by adopting follow-up procedures that are intended to counteract the effects of variance heterogeneity. The contributions by Welch (1947) and Hochberg (1976) are applicable. The former technique uses the sample variances and sizes to obtain a modified CV, whereas by adopting Hochberg's work one uses a follow-up test statistic that has a nonpooled estimate of the standard error of the mean difference.

The present study, therefore, compared the rates of Type I error for nine LSD two-stage procedures to empirically verify in particular that the ANOVA F followed by Student's t sequence does not provide control of FWI and to find an improved two-stage strategy. To provide recommendations for controlling the overall rate of Type I error for a set of pairwise comparisons, we compare our results with simultaneous multiple comparison procedures.

Definition of Test Statistics

Let X_{ik} represent the ith observation in the kth group, where $i = 1, \ldots, n_k$, $k = 1, \ldots, K$ and $\sum n_k = N$. The X_{ik}s are independent normal variates with expected values μ_k and variances σ_k^2. The best linear unbiased estimates of μ_k, σ_k^2 and σ^2 are

$$\bar{X}_{.k} = \sum_i X_{ik}/n_k,$$

$$s_k^2 = \sum_i (X_{ik} - \bar{X}_{.k})^2/(n_k - 1),$$

$$MSW = \sum_{ik} (X_{ik} - \bar{X}_{.k})^2/(N - K),$$

respectively (MSW = mean square within groups).

The omnibus test statistics are as follows: For ANOVA F,

$$F = \frac{\sum_k n_k(X_{.k} - \bar{X}_{..})^2/(K-1)}{\sum_{ik} (X_{ik} - \bar{X}_{.k})^2/(N-K)},$$

where $\bar{X}_{..} = \sum_k n_k\bar{X}_{.k}/N$. When the population variances are homogeneous, F is distributed as a F variable with $K-1$ and $N-K$ degrees of freedom. For Brown and Forsythe's (1974) F^*,

$$F^* = \frac{\sum_k n_k(\bar{X}_{.k} - \bar{X}_{..})^2}{\sum_k (1 - n_k/N)s_k^2},$$

where F^* is approximately distributed as F with $K-1$ and f degrees of freedom and f is obtained with the Satterthwaite (1941) approximation,

$$1/f = \sum_k c_k^2/(n_k - 1);$$

$$c_k = (1 - n_k/N)s_k^2/[\sum_k (1 - n_k/N)s_k^2].$$

For Welch's (1951) F'',

$$F'' = [\sum^K w_k (\bar{X}_{.k} - \bar{\bar{X}}_{.})^2/(K-1)]/$$
$$1 + [2(K-2)/(K^2-1)]$$
$$\times \{\sum^K [1/(n_k - 1)](1 - w_k/\sum w_k)^2\},$$

where $w_k = n_k/s_k^2$ and

$$\bar{\bar{X}}. = \sum^{K} w_k \bar{X}._k / \sum^{K} w_k.$$

The Welch statistic is approximately distributed as an F variable with $K-1$ and

$$\nu_w = (K^2 - 1)/\{3 \sum [1/(n_k - 1)] \times (1 - w_k/\sum w_k)^2\}$$

degrees of freedom.

The pairwise comparisons can be assessed for statistical significance with the statistic

$$t = (\bar{X}._k - \bar{X}._{k'})/[\text{est}\ (\sigma_k^2/n_k + \sigma_{k'}^2/n_{k'})]^{\frac{1}{2}}$$

to test the null hypothesis, $\mu_k - \mu_{k'} = 0$ $(k \neq k')$. The t designation is not meant to imply that this statistic fits Student's t distribution, particularly under heterogeneous variances.

The estimated standard errors of the mean differences for the pairwise tests are (a) Student's pooled denominator,

$$(MSW/n_k + MSW/n_{k'})^{\frac{1}{2}};$$

(b) Hochberg's (1976) nonpooled denominator,

$$[2 \max (s_k^2/n_k, s_{k'}^2/n_{k'})]^{\frac{1}{2}};$$

and (c) the Behrens (1929)–Fisher (1935) denominator (used with a modified CV),

$$(s_k^2/n_k + s_{k'}^2/n_{k'})^{\frac{1}{2}}.$$

The three forms of t use a $t_{(\alpha/2)}$ CV. The error df for Student's and Hochberg's ts equals $N - K$. The variable error $df(\nu_w)$ for the third denominator is Welch's (1947) solution for ν_w, where

$$\nu_w = \frac{(s_k^2/n_k + s_{k'}^2/n_{k'})^2}{\dfrac{(s_k^2/n_k)^2}{n_k - 1} + \dfrac{(s_{k'}^2/n_{k'})^2}{n_{k'} - 1}}.$$

Methods of the Simulation Study

Pseudorandom normal observations of sizes 29, 41, 65, and 89 were obtained from the Marsaglia, MacLaren, and Bray (1964) random number generator.[1] An omnibus test statistic was then computed on the data. If the observed value of the omnibus test exceeded a 5% critical value based on 3 and error df the FORTRAN program was used to compute the six pairwise comparisons. This was repeated until 1,000 different significant results had been obtained on the omnibus tests. The average of the six per-comparison rates of Type I error (the average PCI) was then obtained by dividing the total number of false rejections by 6,000 (6 comparisons × 1,000 simulations).

If the observed value of the omnibus test did not exceed its critical value, the computer program was set up to return to the random number generator. For normally distributed observations with means of zero and a common variance of one, approximately 20,000 calls to the generator are necessary to obtain 1,000 significant omnibus tests for a 5% level of significance. To optimize programming efficiency, each omnibus-to-pairwise-comparisons combination was run separately. However, the starting numbers for the random number generator were kept the same for each combination, and consequently each of the omnibus tests started with the identical set of data.

A heterogeneous variance condition was also investigated. The unequal variances (.104, .790, .810, and 2.296) were inversely paired with the unequal sample sizes (i.e., smallest σ_k^2 with largest n_k and largest σ_k^2 with smallest n_k). This particular type of pairing was chosen because it delineates the case in which the rates of Type I error are inflated in tests using *MSW*.

Results and Discussion

Table 1 presents the results of the simulation study. The FWI values associated with each omnibus test are the probabilities that the omnibus test will yield a significant result and that the pairwise tests using PCI CVs will result in at least one significant pair. Since the latter usually happens whenever the omnibus test is significant, the FWI values in each of these columns are basically the same. Clearly, the major determinants of these FWI values are the characteristics of the omnibus tests themselves. The average PCIs are the average values of the per-comparison rates of Type I errors on the six comparisons of each experiment, given that a significant omnibus test was obtained; this is a conditional probability.

Looking at the top set of values in Table 1 under the homogeneous variance condition, one sees that all of the tests are very similar and perform about as expected when this major assumption has been met. The PCI averages are higher than .05 because they are conditional probabilities (which were computed here only when the sample means were sufficiently divergent to yield a significant omnibus test). In summary, the values in the upper section of the table suggest that any of

[1] See Golder and Settle (1976) and Payne (1977) for a description and evaluation of the Marsaglia, MacLaren, and Bray (1964) random number generator.

Table 1
Type I Error Rates for Various Two-Stage Procedures

Procedure used on pairs	ANOVA F Average PCI	ANOVA F FWI[a]	Brown and Forsythe's (1974) F^* Average PCI	Brown and Forsythe's (1974) F^* FWI[b]	Welch's (1951) F'' PCI	Welch's (1951) F'' FWI[c]
	$n_1 \neq n_2 \neq n_3 \neq n_4$; $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$					
Student's pooled t	.379	.049	.375	.049	.367	.048
Welch's (1947) t''	.365	.049	.370	.049	.372	.048
Hochberg's (1976) t	.288	.047	.295	.048	.299	.047
	$n_1 < n_2 < n_3 < n_4$; $\sigma_1^2 > \sigma_2^2 > \sigma_3^2 > \sigma_4^2$					
Student's pooled t	.460	.177[d]	.521	.064[e]	.462	.049[f]
Welch's (1947) t''	.214	.119	.396	.064	.380	.050
Hochberg's (1976) t	.072	.052	.141	.033	.211	.039

Note. PCI = per-comparison Type I error rate; FWI = familywise risk of Type I error.
[a] FWIs based on 20,476 simulations. [b] FWIs based on 20,508 simulations. [c] FWIs based on 20,912 simulations. [d] FWIs based on 5,660 simulations. [e] FWIs based on 15,487 simulations. [f] FWIs based on 19,896 simulations.

the omnibus tests can be used as the first step in the two-stage procedure when one has homogeneous variances.

The values in the lower section of Table 1 indicate that the ANOVA F is not to be trusted as the first stage when variances are heterogeneous and sample sizes are unequal. The FWI values are substantially above .05, except when the ANOVA is followed by Hochberg's (1976) conservative test. Probably the most commonly used technique is to calculate an ANOVA F as the first step and then follow this by Student's t tests if the F is significant. It is disturbing to find that under the heterogeneous variances condition with nominal alphas of .05, the FWI value is about .18 and that given a significant F, the mean probability of a Type I error in the comparisons that follow is .46.

For the unequal sample size–unequal variance case, the other two omnibus tests provide reasonable control of FWI, with the Welch (1951) F'' being slightly superior to the Brown and Forsythe (1974) F^*. However, although either is an improvement over the ANOVA F, it would be a mistake to follow either of these omnibus tests with a follow-up test that made use of MSW, since the standard errors based on this value would be inappropriate for various pairwise comparisons of the means (Games & Howell, 1976, p. 119). Thus, the choice for a follow-up test is limited to the Welch t'' (1947) or the Hochberg (1976) t procedures. All LSDs using the Hochberg t effectively controlled the rate of Type I error. However, the rates were conservative when the omnibus test used was F^* or F''. Because of this conservativeness we would expect these two LSDs to be relatively less powerful. The Welch F''–Welch t'' LSD also provided Type I control, but was not conservative and consequently should be relatively more powerful. However, the F^*–Welch t'' LSD proved to be slightly liberal and can be discounted if the user wants to maintain the number of Type I errors at or below the significance level.

Though not presented here, Type I error rates per family were collected to compare the Welch LSD(F''–t'') and simultaneous multiple comparison approaches.[2] Interestingly, the per-family rates were not very disparate. Consequently, though our preference is for the simultaneous multiple comparison

[2] The per-family Type I error rate is equal to the number of Type I errors made on the 6,000 comparisons (after a significant omnibus test) divided by the total number of families, or, here, by experiments run in the simulation (Miller, 1966, p. 5). The rates for the multiple comparison procedures (Tukey's, Note 1, procedure using the Welch, 1947, CV as suggested by Games & Howell, 1976) were obtained in another simulation study.

approach in testing pairwise comparisons, the data indicate that some two-stage least significant difference procedures can indeed provide Type I error control.

Reference Note

1. Tukey, J. W. *The problem of multiple comparisons.* Unpublished manuscript, Princeton University, 1953.

References

Behrens, W. V. Ein Beitrag zur Fehlerberechung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher*, 1929, *68*, 807–837.

Boneau, C. A. The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 1960, *57*, 49–64.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problem: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 1954, *25*, 290–302.

Brown, M. B., & Forsythe, A. B. The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 1974, *16*, 129–132.

Carmer, S. G., & Swanson, M. R. An evaluation of ten multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association*, 1973, *68*, 66–74.

Fisher, R. A. The fiducial argument in statistical inference. *Annals of Eugenics*, 1935, *6*, 391–398.

Fisher, R. A. Design of experiments (5th ed.). Edinburgh, Scotland: Oliver & Boyd, 1949.

Games, P. A., & Howell, J. F. Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1976, *1*, 113–125.

Golder, E. R., & Settle, J. C. The Box–Muller method for generating pseudo-random normal deviates. *Applied Statistics*, 1976, *25*, 12–20.

Hochberg, Y. A modification of the T-method of multiple comparisons for a one-way layout with unequal variances. *Journal of the American Statistical Association*, 1976, *71*, 200–203.

Kohr, R. L., & Games, P. A. Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *Journal of Experimental Education*, 1974, *43*(1), 61–69.

Marsaglia, G., MacLaren, M. D., & Bray, T. A. A fast procedure for generating normal random variables. *Communications of the ACM*, 1964, *7*, 4–10.

Miller, R. G., Jr. *Simultaneous statistical inference.* New York: McGraw–Hill, 1966.

Payne, W. H. Normal random numbers: Using machine analysis to choose the best algorithm. *ACM Transactions on Mathematical Software*, 1977, 3(4), 346–358.

Satterthwaite, F. E. Synthesis of variance. *Psychometrika*, 1941, *6*, 309–316.

Welch, B. L. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 1947, *34*, 28–35.

Welch, B. L. On the comparison of several mean values: An alternative approach. *Biometrika*, 1951, *38*, 330–336.

Received April 16, 1978 ■

CONTENTS (continued)

VOLUME 86 / NUMBER 5 SEPTEMBER 1979

Psychological Bulletin

(Continued on inside back cover)

UBLISHED BIMONTHLY BY THE AMERICAN PSYCHOLOGICAL ASSOCIATION
ISSN 0033-2909

R. J. Herrnstein, *Editor, Harvard University*

Gene V Glass, *Associate Editor, University of Colorado*

Susan Herrnstein, *Assistant to the Editor*

The *Psychological Bulletin* publishes evaluative reviews and interpretations of substantive and methodological issues in the psychological research literature. The Journal reports original research only when it illustrates some methodological problem or issue. Discussions of methodological issues should be aimed at the solution of some particular research problem on psychology, but should be of sufficient breadth to interest a wide readership among psychologists; articles of a more specialized nature can be directed to the various statistical, psychometric, and methodological journals. The *Bulletin* does not publish original theoretical articles; these should be submitted to the *Psychological Review*.

Abstracts: All articles must be preceded by an abstract of 100–175 words. Detailed instructions for preparation of abstracts appear in the *Publication Manual of the American Psychological Association* (2nd ed.), or they may be obtained from the Editor or from APA Central Office.

Blind review: Because reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the author's identity or affiliation should be on separate pages. Every effort should be made to see that the manuscript itself contains no clues to the author's identity.

Manuscripts: Submit manuscripts in triplicate to the Editor, R. J. Herrnstein, *Psychological Bulletin,* Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, according to instructions provided below.

Instructions to Authors: Authors should follow the directions given in the *Publication Manual of the American Psychological Association* (2nd ed.). Instructions on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors are requested to refer to the "Guidelines for Nonsexist Language in APA Journals" (Publication Manual Change Sheet 2, *American Psychologist,* June 1977, pp. 487–494) before submitting manuscripts to this journal. All manuscripts should be submitted in triplicate and all copies should be clearly readable, and on paper of good quality. Dittoed copies are not acceptable and will not be considered. Authors are cautioned to carefully check the typing of the final copy and to retain a copy of the manuscript to guard against loss in the mail.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1979: $40 domestic, $42 foreign, $7 single issue. APA member rate: $15. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Dept., APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in January) in one volume per year by the American Psychological Association, Inc., 1400 North Uhle Street, Arlington, Va. 22201. Printed in the U.S.A. Second-class postage paid at Arlington, Va., and at additional mailing offices.

Vol. 86, No. 5 September 1979

Psychological Bulletin

Human Spatial Abilities: Psychometric Studies and Environmental, Genetic, Hormonal, and Neurological Influences

Mark G. McGee
Texas A & M University

The spatial abilities literature is reviewed. Psychometric consideration encompasses both factor analytic studies that conclusively demonstrate the existence of at least two Spatial factors—Visualization and Orientation—and predictive validity studies that argue for these factors' social relevance. Sex differences in various aspects of perceptual-cognitive functioning (e.g., mathematics, field independence) are interpreted as a secondary consequence of differences with respect to spatial visualization and spatial orientation abilities. Sources of variation in performance on spatial tests including environmental, genetic, hormonal, and neurological are considered, with special emphasis on age and sex differences. Evidence that variation in spatial test scores is to some degree heritable remains positive; however, the X-linked recessive gene hypothesis that has served as a tentative explanation for sex differences in spatial abilities and for the mode of genetic transmission is not supported strongly in recent studies. Neurological studies showing variations in the lateral organization of the human brain provide experimental evidence for a structural source of the variation in spatial abilities, and this evidence is reviewed as it relates to human handedness and cerebral bilateralization for spatial and linguistic functions.

The purpose of this article is threefold: (a) to summarize psychometric studies of human spatial abilities, (b) to examine the consistencies and disagreements in relation to the hypothesis that sex differences in various aspects of perceptual-cognitive functioning (e.g., mathematics, field dependence–field independence) are a secondary consequence of differences with respect to spatial visualization and spatial orientation abilities, and (c) to review the literature with reference to environmental, genetic, hormonal, and neurological influences that interact in producing individual variation in spatial test scores.

Psychometric Studies

Early Factor Analytic Studies

Historically, the identification of the Spatial factor has roots in the study of mechani-

This work was supported in part by funds from the Office of University Research, Texas A & M University.

The author wishes to thank Thomas J. Bouchard, Jr., Marvin D. Dunnette, Irving I. Gottesman, Rene Dawis, Herbert L. Pick, Sharon Stamos, David Lundberg, and Sheri A. Berenbaum for their helpful comments, and he also thanks the anonymous reviewers of an earlier version of this article. Thanks are due also to Sandra Scarr and Richard Weinberg for the use of computer programs developed under National Institute of Child Health and Human Development Grant HD-08016.

Requests for reprints should be sent to Mark G. McGee, Department of Psychology, Texas A & M University, College Station, Texas 77843.

 0033-2909/79/8605-0889$00.75

cal aptitude (Cox, 1928; Paterson, Elliott, Anderson, Toops, & Heidbreder, 1930; Smith, 1964; Stenquist, 1922) and practical ability (W. P. Alexander, 1935; Kohs, 1923; McFarlane, 1925). In one of the earliest studies of "practical ability," McFarlane found evidence for a group factor over and above general intelligence (*g*). She described individuals in possession of the practical ability underlying this factor as being adept at judging concrete spatial relations.

Since 1925, numerous factor analytical studies have yielded a Spatial factor mathematically distinct from verbal ability. Kelley (1928) identified a Spatial factor and described it as the mental manipulation of shapes. Brown and Stephenson (1933) found that two tests in particular had substantial loadings on their Spatial factor: fitting shapes, which is a paper-form-board test, and a test of dot perception. Koussy (1935) identified a group factor (K) among 28 tests administered, and he concluded from introspective reports by participants in the study that the mental processes active in the solution of problems involving the K factor were characterized by the "ability to obtain and the facility to utilize, spatial imagery" (p. 86). A similar conclusion was reached independently by Smith (1938). Murphy (1936) factor analyzed scores from numerous verbal, nonverbal, and mechanical tests and concluded that mechanical ability included two factors: Speed of Eye–Hand Coordination and Mental Manipulation of Spatial Relations. Clarke (1936) reported the Spatial factor loadings of spatial and verbal tests to be inversely related among girls ranging in age from 12 to 15 years, a relationship that has been replicated with males as well as females (Andrew, 1937a, 1937b; Broverman & Klaiber, 1969; Emmett, 1949; Estes, 1942; Heston, 1943; Morris, 1939; Slater, 1940; Smith, 1938; Swineford, 1948; Thurstone, 1944, 1947; Wittenborn, 1945). This relationship is less likely to obtain for spatial-performance tests highly correlated with general intelligence (e.g., the Block Design subtest of the Wechsler IQ scales, or the Kohs Block Design Test that correlates .80 with Binet IQ scores, Kohs, 1923).

Recent Factor Analytic Studies

Several recent studies have discussed the presence of a Spatial factor among batteries of tests administered (DeFries et al., 1974; DeFries et al., 1976; Goldberg & Meredith, 1975; Gough & Olton, 1972; Hakstian & Cattell, 1974; Yen, 1975; Zonderman, Vandenberg, Spuhler, & Fain, 1977), and others have shown a Spatial factor that reliably appears in cognitive test scores of different racial, ethnic, and socioeconomic (SES) groups. Michael (1949), for instance, found a generally similar factor structure among ability measures on black and white Air Force cadets. Flaugher and Rock (1972) found similar factor structures in a battery of 9 cognitive measures administered to black, white, Mexican American, and Oriental American high school males. Among Americans of European and Japanese ancestry, DeFries et al. (1974) found four factors in the 15 cognitive tests administered: Spatial Visualization, Verbal, Memory, and Perceptual Speed. Humphreys and Taber (1973) found six factors in 21 ability measures from the Project TALENT test battery, and the factor structure was similar for 9th-grade boys in the top and bottom quartiles on SES. Backman (1971, 1972) compared 12th-grade Jewish and non-Jewish whites and Oriental Americans in performance on six mental ability factors—Spatial, Verbal, English Language, Math, Perceptual Speed, and Memory—and found ethnicity and SES to have much less influence on the ability profile than did sex. Similar subtest intercorrelation patterns have been reported by Nichols (1971) and Scarr-Salapatek (1971) for U.S. whites and blacks. These studies all agree with the suggestion "that underlying dimensions of ability vary little if at all across U.S. racial-ethnic groups" (Loehlin, Lindzey, & Spuhler, 1975, p. 179).

An equally important emphasis of recent factor analytic research has been that of disentangling various subabilities that characterize the Spatial factor. The available evidence conclusively demonstrates the existence of at least two Spatial factors: Visualization and Orientation. Table 1 presents a summary of Spatial Visualization and Spatial

Orientation factor symbols and descriptions. Although factor names and symbols differ in the four studies cited in Table 1, factor descriptions are strikingly similar.

The first clear evidence for the existence of spatial abilities resulted from an impressive series of factor analytic studies initiated by L. G. Humphreys of the Army Air Force

Table 1
Summary of Spatial Visualization and Spatial Orientation Factor Symbols and Descriptions

	Spatial visualization factor		Spatial orientation factor	
Investigator	Symbol	Description	Symbol	Description
Guilford and Lacey, (1947)	V_z	An ability to imagine the rotation of depicted objects, the folding or unfolding of flat patterns, the relative changes of position of objects in space, the motion of machinery. This visualization factor is strongest in tests that present a stimulus pictorially and in which some manipulation or transformation to another visual arrangement is involved.	SR	An ability to determine relationships between different spatially arranged stimuli and responses and the comprehension of the arrangement of elements within a visual stimulus pattern.
Thurstone (Note 1)	S_2	An ability to visualize a configuration in which there is movement or displacement among the internal parts of the configuration.	S_1	An ability to recognize the identity of an object when it is seen from different angles or an ability to visualize a rigid configuration when it is moved into different positions.
			S_3	An ability to think about those spatial relations in which the body orientation of the observer is an essential part of the problem.
French (1951)	V_i	An ability to comprehend imaginary movements in three-dimensional space or the ability to manipulate objects in the imagination.	S	An ability to perceive spatial patterns accurately and to compare them with each other.
			SO	An ability to remain unconfused by the varying orientations in which a spatial pattern may be presented. Dimensionality is less important to the factor than the rotational position of presentations.
Ekstrom, French, and Harman (Note 3)	VZ	An ability to manipulate or transform the image of spatial patterns into other arrangements; requires either the mental restructuring of a figure into components for manipulation or the mental rotation of a spatial configuration in short term memory, and it requires performance of serial operations, perhaps involving an analytic strategy.	S	An ability to perceive spatial patterns or to maintain orientation with respect to objects in space; requires that a figure be perceived as a whole.

Note. Adapted from Michael, Guilford, Fruchter, and Zimmerman (1957, Table 1, p. 188).

(AAF) (Guilford & Lacey, 1947; Guilford & Zimmerman, 1947). The results, based on repeated analyses from AAF tests administered to thousands of military personnel, indicated two Spatial factors: Spatial Visualization (V_z) and Spatial Relations (SR). Visualization was described as the ability to imagine the rotation of depicted objects, the folding or unfolding of flat patterns, the relative changes of position of objects in space, or the motion of machinery. Spatial Relations was described as comprehension of the arrangement of elements within a visual stimulus pattern.

Thurstone (1938) had isolated a Space factor that he designated as a facility in spatial and visual imagery, but it was not until 1950, when he found several primary abilities in visual thinking and related three of these to visual orientation in space, that the differentiation between them became better understood. The first of these three factors was designated S_1 (Thurstone, Note 1). He asserted that it represented the ability to recognize the identity of an object when it was seen from different angles. This factor was also characterized by the ability to visualize a rigid configuration when it was moved into different positions. The second factor identified by Thurstone (Note 1), S_2, was said to represent the ability to imagine movement or internal displacement among the parts of a total configuration. Thurstone's distinction between abilities to imagine transformation of wholes (S_1) versus parts (S_2) is unique and needs verification. The third factor (S_3) was identified by Thurstone as the ability to think about those spatial relations in which the body orientation of the observer is an essential part of the problem.

In 1951, French identified a Visualization factor (V_i), described as the ability to mentally manipulate three-dimensional objects, and an Orientation factor (SO), described as an ability to remain unconfused by the varying orientations in which a spatial pattern may be presented.

French, Ekstrom, and Price (Note 2) described V_z as an ability to manipulate or transform the image of spatial patterns into other visual arrangements and spatial orientation as an ability to perceive spatial patterns or maintain orientation with respect to objects in space. More recently, Ekstrom, French, and Harman (Note 3), in their manual for Kit of Factor-Referenced Cognitive Tests, have suggested that visualization ability requires that a figure be mentally restructured into components for manipulation, whereas the whole figure is manipulated in spatial orientation. Both spatial orientation and visualization require short term visual memory. Orientation requires only mental rotation of the configuration; however, visualization, according to this more recent definition, requires both rotation and the performance of serial operations. The requirement that a figure is perceived as a whole in spatial orientation but must be mentally restructured into components for manipulation in visualization (Ekstrom et al., Note 3) is reminiscent of Thurstone's (Note 1) distinction between abilities to imagine transformation of wholes (S_1) and parts (S_2).

Corroborative evidence for the existence of at least two Spatial factors has been provided. Guilford, Fruchter, and Zimmerman (1952), for example, found 12 factors in a test battery of 46 tests (including a Spatial Relations and a Visualization factor). The Visualization factor was represented as an ability to mentally manipulate elements of a pattern, and the Orientation factor was represented as an ability to determine spatial orientation with respect to one's body.

Summary

Factor analytic studies of the Spatial factor began with the study of practical and mechanical ability during the mid-1920s. Some investigators have found evidence for distinct factors of practical and mechanical ability, in addition to a Spatial factor; others have refuted this distinction (Dempster, 1948; Leff, 1949; Price, 1940; Slater, 1940; Watts, 1953; Williams, 1948). Although the debate over the existence versus nonexistence of a Spatial factor characterizes much of the literature prior to 1930, a plethora of factor studies since that date have provided strong and consistent support for the existence of two distinct spatial abilities: visualization and orientation.

As for visualization, the S_2 factor proposed by Thurstone (Note 1) is similar to the V_i factor proposed by French (1951), the V_z factor reported by the AAF researchers (Guilford & Lacey, 1947), and the V_z factor described by Ekstrom et al. (Note 3). All involve the ability to mentally manipulate, rotate, twist, or invert a pictorially presented stimulus object. The underlying ability seems to involve a process of recognition, retention, and recall of a configuration in which there is movement among the internal parts of the configuration (S_2) or the recognition, retention, and recall of an object manipulated in three-dimensional space (V_i) or which involves the folding or unfolding of flat patterns (V_z). The Spatial Visualization Test of the French et al. (Note 2) Kit of Reference Tests of Cognitive Factors, for example, requires the examinee to mentally fold and unfold a piece of paper and choose the alternative that represents the paper after it has been unfolded. The Guilford-Zimmerman (1953) Visualization Test consists of a picture of an alarm clock and a sphere with directional arrows, and the examinee is required to visualize the rotation of the clock as it is moved into different positions according to the directions of the arrows.

As for spatial orientation, the S_1 and S_3 factors proposed by Thurstone (Note 1), the SO factor proposed by French (1951), the SR factor proposed by Guilford and Lacey (1947), and the SO and S factors described by Ekstrom et al. (Note 3) are similar. All involve the comprehension of the arrangement of elements within a visual stimulus pattern and the aptitude to remain unconfused by the changing orientation in which a spatial configuration may be presented. The Spatial Orientation Test of the French et al. (Note 2) Kit of Reference Tests of Cognitive Factors requires the examinee to compare cubical blocks and indicate whether they are the same or different according to symbols written on their faces. The Guilford-Zimmerman (1953) Spatial Orientation Test requires the examinee to imagine riding in a boat whose prow is always visible in the foreground of the pictures that comprise each item and to choose among the alternative new directions of the boat.

Regarding factor analytic studies of spatial abilities in general, some qualifications need to be made. First, after 70 years of psychometric research, there is still vast disagreement about just how best to classify standard tests of spatial abilities. Further factor analytic studies are indicated. Second, the influence of test-item difficulty on factor structure needs to be further explored. Myers' (Note 4) suggestion that visualization test items are usually more difficult than orientation items has been supported (Zimmerman, 1954a, 1954b), although not replicated. Third, spatial tests consisting of both two- and three-dimensional items are used with equal frequency, but little is known about how dimensionality contributes to the factor structure of spatial tests. Fourth, studies showing positive correlations between tests of spatial visualization and orientation (Borich & Bauman, 1972; Goldberg & Meredith, 1975; Karlins, Schuerkoff, & Kaplan, 1969; Roff, 1952; Yen, 1975) illustrate the need for further factor analytic research to clarify the specificity of the Visualization and Orientation factors.

Predictive Validity Studies

Information from predictive validity studies reinforces our judgment concerning the practical use of visualization and orientation abilities and proves their social significance. Historically, predictive validity has been important in the selection of candidates to fill openings in industry, colleges, and the armed forces. Predictive validity is at issue when the purpose of an instrument is to estimate some important form of behavior (Nunnally, 1978). The degree of relationship between an instrument and its criterion indicates the instrument's predictive validity. The use of tests to predict success in school is a widely accepted practice, and the educational institution beyond the elementary school level that does not use a standardized testing procedure in its evaluation of incoming students is now the exception.

To what practical use can we direct our knowledge of human spatial abilities? What are the implications of this knowledge for use in everyday life? And to what extent is

this knowledge useful in predicting socially significant and relevant behavior? We have two lines of evidence directly relevant to these issues: first, the literature regarding the value of spatial tests for selecting workers for industrial jobs and predicting job performance and, second, the literature regarding the use of spatial tests for the prediction of success in vocational–technical training programs.

Selecting workers for industrial jobs and predicting job performance. The value of spatial tests for use in making personnel selection decisions has been well documented elsewhere. Ghiselli (1966, 1973), for example, has summarized much of the literature relevant to the predictive validity question of occupational aptitude tests. The U.S. Employment Service (1957) has listed those occupations requiring a high level of spatial ability. Four job categories—engineering, science, drafting, and designing—account for nearly 85% of all jobs listed.

Predicting success in vocational–technical training programs. The earliest evidence for the prediction of success in school work with a test specifically designed to measure spatial ability was provided by O'Connor (cited in Smith, 1964). Using the O'Connor Wiggly Block Test, he found predictive validities of .62 and .42 for shop grades in two groups of vocational school boys.

Paterson et al. (1930), in their massive investigation of mechanical ability, found spatial tests to be especially useful in predicting success in various junior high school and technical school courses. The spatial test battery (consisting of several tests including the Minnesota Paper Form Board Test, the Link Spatial Relations, and the Packing Blocks Tests) yielded a multiple correlation of .60 with success in shop courses and correlated .07 with intelligence.

Holliday (1940) administered a battery of spatial tests to several groups of trade school apprentices, engineer apprentices, and shop students. The verbal test administered correlated .07 with proficiency in technical drawing, whereas the corresponding correlation of the spatial battery was .66. In a subsequent investigation, Holliday (1943) administered a series of spatial, mechanical, and verbal tests to several groups of toolmakers and engineering and trade apprentices. Again it was found that the spatial tests yielded higher correlations with mechanical drawing than did the verbal tests administered. Slater (1941) conducted a validity study in which he used seven spatial tests in addition to a test of verbal ability to predict criterion estimates of engineering drawing ability and general apprenticeship ability. Drawing ability correlated .41 with composite scores on the spatial tests but only .26 with verbal ability. Shuttleworth (1942) found that tests of spatial ability showed higher correlations with grades obtained in junior technical school than tests of verbal and mechanical ability. Although the correlation between grades and the verbal test administered was only .20, the corresponding correlation for the Space Perception, Memory for Design, and Form Relations Tests were .46, .45, and .44, respectively. Hunter (1945) used the Minnesota Paper Form Board and the Otis Intelligence Test (Form B) among other tests in predicting course success for sophomore and junior machine shop students. Correlations against this criterion were .45 for the spatial test and .28 for the intelligence test. Smith (1948) administered 11 tests to first- and second-year pupils in a Scottish secondary school. The spatial battery included tests of area discrimination, completion, fitting shapes, form equations, classification, form-figure analogies, form recognition, pattern perception, and drawing. The Otis Intelligence Test (Form B) was also used. The spatial test battery was predictive of success in engineering drawing ($r = .66$) and art ($r = .39$). Correlations between grades in these two subjects and the Otis Intelligence Test were $-.07$ and .19, respectively.

Two validity studies reported by Smith (1964), corroborating the results of those presented above, were conducted by Knight (1949) and by Oxlade (1951). Knight reported results from a follow-up study of Middlesex Junior Technical School entrance examinees to determine the validity of spatial tests in predicting successful completion of the program. The battery of tests administered included a form relations test, a

memory-for-designs test, and a space perception test. Headmaster's ratings were also obtained. The criteria included performance in the program at the end of the first and second years. Correlations between predictor and criterion variables demonstrated the superiority of spatial tests over headmaster's ratings for selecting candidates. In the first year, the predictive value of headmaster's ratings and selection tests was similar, but, as the course became more advanced and technical in its second year, the predictive value of the selection tests became greater, whereas that of the headmaster's ratings decreased. Martin (1951) found that a spatial relations test predicted shop achievement and classroom grades among 45 auto mechanics in a California technical school. And in a related study, Hunt (1973) demonstrated the usefulness of a spatial test in predicting course achievement in a computational science curriculum.

Smith (1964) reported a follow-up study of pupils selected for technical education in which the major objective was to assess the predictive validity of various selection tests, including tests of spatial and verbal ability, against criteria of success in technical courses in a secondary technical school. Data were collected on students in a number of technical courses at 3 years after and at 5 years after the original selection examination. Both spatial tests that were used showed substantial correlations with criterion examination scores in all technical courses. The most highly significant regression coefficients for the two spatial tests tended to be in the areas of metalwork, woodwork, handicraft, and drawing (geometrical, building, and engineering). In contrast, the verbal reasoning test that was used showed low validity for the prediction of success in these same areas. Similar results had been provided from two previously conducted validity studies. Holzinger and Swineford (1946) examined the predictive validity of a battery of spatial tests with grades in various school courses and found that test scores correlated .002 with English, −.003 with biology, and −.06 with foreign language but .23 with plane geometry, .46 with shops and crafts, and .69 with drawing. The spatial tests included in the battery were of visual imagery, cubes, punched holes, figures, form relations, pattern perception, and drawings.

In related studies the predictive validity of spatial orientation and spatial visualization tests has been demonstrated. Hills (1957), for example, reported a validation study on the relationship between several measures of aptitude and success in college mathematics. The Spatial Visualization and Spatial Orientation Tests (Guilford & Lacey, 1947) were among the best predictors of mathematical grades, showing higher validity coefficients than the verbal tests used. Karlins et al. (1969) investigated cognitive factors relating to architectual creativity among graduating achitecture students and found a significant correlation of .49 between scores on Thurstone's Cubes Test (a measure of spatial orientation, Thurstone, Note 1) and quality of independent work completed. And the space subtest of the Differential Aptitude Test battery (Bennett, Seashore, & Wesman, 1974) is predictive of drafting ($r = .42$), shop mechanics ($r = .47$), and watch repair ($r = .69$), none of which is well predicted by verbal tests.

Relationship Between Spatial Abilities and Various Perceptual-Cognitive Tasks

Differential psychologists have been interested in spatial visualization and orientation because male superiority on tasks requiring these abilities is among the most persistent of individual differences in all the abilities literature (Anastasi, 1958; Buffery & Gray, 1972; Garai & Scheinfeld, 1968; Harris, 1978; Maccoby & Jacklin, 1974; McGee, 1977; O'Connor, 1943; Sherman, 1971; Smith, 1964; Tyler, 1965).

The widely documented sex difference on tests of spatial visualization and spatial orientation as well as on numerous tasks requiring these abilities does not reliably appear until puberty (Drew, 1944; Emmett, 1949; Fruchter, 1954; Gardner, Jackson, & Messick, 1960; Harris, 1978; Maccoby, 1966; Slater, 1941; Witkin et al., 1954). However, in studies in which differences have been reported in younger samples, boys typically showed superiority in performance (for

reviews, see Harris, 1978; Maccoby & Jacklin, 1974; Smith, 1964), which is particularly puzzling in light of the general maturational advantage in physical and cognitive development enjoyed by girls (Garai & Scheinfeld, 1968; Maccoby & Jacklin, 1974; Money & Ehrhardt, 1972; Waber, 1976, 1977).

In this section we examine the consistencies and disagreements in relation to the hypothesis that sex differences in various aspects of perceptual-cognitive functioning are a secondary consequence of differences with respect to spatial visualization and spatial orientation abilities.

Visualization Ability

Recall that spatial visualization involves the ability to mentally rotate, manipulate, and twist two- and three-dimensional stimulus objects. It is tempting to consider visualization as fundamental to good performance in various areas of mental functioning. Two areas are considered below: imagery and mathematical ability.

Imagery. Galton's (1880, 1883) pioneering work on individual differences in imagery led the way for researchers who have since been intermittently interested in the issue. The study of mental transformations of spatial images is returning to popularity in psychology and represents an attempt to examine the "major representational alternative" to language (Neimark & Santa, 1975) to more fully understand the mental processes involved in the solution of tasks that are difficult to solve by verbally mediated processes. For reviews, see Neimark & Santa, 1975; Pylyshyn, 1973; Lane, Note 5).

The empirical study of imagery has emphasized the measurement of imagery vividness and imagery control (Betts, 1909; Costello, 1957; Galton, 1880, 1883; Gordon, 1949; Marks, 1972; Richardson, 1969, 1972; Sheehan, 1967a, 1967b). It has been suggested (Richardson, 1972) that spatial manipulation involves imagery control and that imagery vividness and control may be related to individual differences in performance on tests of spatial ability. Those who lack control over their imagery have been described by Galton (1883) as having "difficulty in shifting their mental view of an object and examining it at pleasure in different positions" (p. 75). As noted by Koussy (1935), solving problems on spatial tests requires mental imagery and the ability to obtain and the facility to use visual spatial imagery. The role played by visual imagery in tests of spatial visualizing ability is unclear, however.

Although no studies have examined the relationship between imagery vividness, imagery control, and performance on tests of spatial ability, the work by Shepard and his associates on mental transformation of visual images has gained considerable attention. They have shown that the reaction time required for performing an instructed transformation is a linearly increasing function of the number of rotations or foldings required for determining whether two differently oriented objects have the same or different shapes (Shepard & Metzler, 1971) and for determining whether arrows on two sides will meet when a flat six-sided figure is folded into a cube (Shepard & Feng, 1972). This research has advanced our understanding of the nature of the underlying mental process involved in the solution of spatial visualization tasks such as those that appear on paper-folding and surface development tests of spatial ability. Further, it suggests the principle of "second-order isomorphism" (Shepard & Chipman, 1970), meaning that the events occurring during "imagining" of an external process (e.g., paper folding) are similar to the events occurring during the "perceiving" of an external process. The orderly relationship between (a) time required to recognize that two-perspective drawings portray objects of the same three-dimensional shape and (b) the angular differences in the portrayed orientations of the two objects (Shepard & Metzler, 1971) implies that while one is in the course of imagining the external process, one passes through an orderly set of internal states of special relation to the successive states of the external process (Shepard & Feng, 1972). Metzler and Shepard (1974) and Cooper and Shepard (1975) have described experi-

ments with college age males and females that confirm their earlier reports. No systematic attempt to examine sex differences has been made, although the slope of the reaction time functions tended to be higher for females than males (Metzler & Shepard, 1974), indicating that males require less time than females to solve Shepard's mental transformation tasks.

Mathematical ability. One definition of mathematical ability has been proposed by Hamley (1935), a mathematician and a psychologist. According to this definition, mathematical ability is a compound of general intelligence, visual imagery, and ability to perceive number and space configurations and to retain such configurations as mental patterns. The extent to which spatial ability enters into mathematical ability is suggested by several validity studies. Hills (1957) investigated the relationship between various aptitude tests and criterion performance in college mathematics. The test battery included two spatial tests—one of visualization and one of orientation—from the Guilford-Zimmerman Aptitude Survey (Guilford & Zimmerman, 1953). Subjects were 148 students in three institutions. The two spatial tests had relatively high correlations with course performance (for visualization, $r = .23$; for orientation, $r = .22$) compared to the verbal and reasoning tests administered ($r = .06$ each), which suggests a higher relative importance of spatial ability than verbal ability in college mathematics. And the space subtest of the Differential Aptitude Test battery (Bennett et al., 1974) is predictive of success in school geometry ($r = .57$) and quantitative thinking ($r = .69$). More recently, Eisenberg and McGinty (1977) have shown that spatial visualization test scores are higher among students in calculus courses than in other college courses.

Corroborative evidence has been provided by Smith (1948) and Werdelin (1961). Male superiority in understanding geometric principles and concepts has been reported (Saad & Storer, 1960), and Smith (1964) has suggested that the sex difference "may be another manifestation of the sex difference in spatial ability, reflecting a greater capacity on the part of boys to perceive, recognize and assimilate patterns within the conceptual structure of mathematics" (p. 123).

Orientation Ability

Recall that spatial orientation involves the comprehension of the arrangement of elements within a visual stimulus pattern, the aptitude to remain unconfused by the changing orientations in which a spatial configuration may be presented, and the ability to determine spatial orientation with respect to one's body. This definition raises the question of whether individual differences in various aspects of perceptual-cognitive functioning are a secondary consequence of differences with respect to spatial orientation ability. Empirical research in four areas is particularly relevant to this question: field dependence – field independence, sense-of-direction, Piagetian, and maze tasks.

Field dependence – field independence. One of the more familiar tasks on which sex differences have consistently been found is that of field dependence – field independence (Witkin, 1950; Witkin, Dyk, Faterson, Goodenough, & Karp, 1962; Witkin et al., 1954).

Two tests—the rod-and-frame test and the Embedded Figures Test—seem to have a strong spatial component. The rod-and-frame test, for instance, requires the examinee to adjust a rod to the vertical position in the absence of cues other than the luminescent square frame that surrounds the rod. The frame position as well as the position of the examinee may be tilted in various orientations. Adult females tend to be more dependent on the field in determining the vertical position of the rod than males (Bogo, Winget, & Gleser, 1970; Corah, 1965; Gross, 1959; Kato, 1965, Morf, Kavanaugh, & McConville, 1971; Okonji, 1969; Saarni, 1973; Schwartz & Karp, 1967; J. Silverman, Bucksbaum, & Stierlin, 1973; Vaught, 1965; Witkin et al., 1962; Witkin, Goodenough, & Karp, 1967), and the sex difference is apparent in adolescents as well (Graves & Koziol, 1971; Keogh & Ryan, 1971; J. Silverman et al., 1973; Canavan, Note 6).

The Embedded Figures Test requires the examinee to view and store in memory a simple geometric form and then to recall the

form by identifying it in a more complex geometric figure. The obvious spatial element in this task may account for the sex difference that is so widely documented for adolescents as well as for adults (e.g., Bieri, Bradburn, & Galinsky, 1958; Bigelow, 1971; Corah, 1965; Goldstein & Chance, 1965; Goodenough & Eagle, 1963; Graves & Koziol, 1971; Keogh & Ryan, 1971; Nash, 1973; Okonji, 1969; Witkin, 1950; Witkin et al., 1954). Sex differences on these tasks are less reliable among children under the age of 11 or 12 years (Maccoby & Jacklin, 1974; Witkin et al., 1954).

Sherman (1967) has provided the major theoretical articulation of the relationship between sex differences in spatial abilities and sex differences in field dependence, arguing that the sex difference in field dependence is an artifact of the sex difference in space perception. A sizable body of literature supports her hypothesis. Correlational studies have consistently demonstrated a strong relationship between tests of spatial orientation and measures of field dependence. Gardner et al. (1960) found correlations of Embedded Figures and rod-and-frame test scores with the Guilford-Zimmerman Spatial Orientation Test of .53 and .35, respectively. Thurstone (1944) reported correlations of .43 and .41 between two forms of the Gottschaldt Figures Test, similar to Witkin's Embedded Figures Test, and the Space Test of the Primary Mental Abilities test battery. This finding has been replicated by Podell and Phillips (1959).

Factor analytic studies indicate that tests of spatial abilities and field dependence–field independence emerge together in a factor similar in description to the Spatial Orientation factor discussed previously (Gardner et al., 1960; Hyde, Geiringer, & Yen, 1975; Podell & Phillips, 1959; Thurstone, 1944) and that sex differences in field dependence are eliminated after removing differences in spatial abilities (Hyde et al., 1975).

The presence of a spatial component in tests of field dependence–field independence seems to be a prerequisite for the appearance of sex differences. Measures of field dependence other than the rod-and-frame and Embedded Figure tests that do not have a spatial component (e.g., the rotator-match brightness constancy task and the body steadiness task) have not shown sex differences (Witkin et al., 1954). In light of the spatial nature of both the rod-and-frame and Embedded Figures tasks, the sex difference is understandable and should be narrowly interpreted, not generalized into an all encompassing statement about cognitive style (Harris, 1978).

Sense of direction. Spatial orientation ability is probably important in tasks requiring sense of direction. Berry's (1966) cross-cultural study comparing Eastern Canadian Eskimos from Baffin Island with members of the Temne tribe in Africa supports this hypothesis. The directional sense among Eskimo males and females fostered by extensive travel in hunting is reflected in higher performance on several tasks requiring spatial ability, including the Morrisby Shapes Test and a test of field dependence–field independence.

Spatial ability may enter into another task of directional sense—map reading. Money, Alexander, and Walker (1965) administered their Road Map Test of Direction Sense to over 1,000 children ranging in age from 7 to 18 years. The task consists of a schematic outline map of several city blocks with a standard route through the streets. The examinee is required to mentally follow the route, indicating verbally the direction of various turns (left or right with reference to point of origin). Males on the average performed significantly better than females and the differences were greatest between boys and girls of older ages.

Piagetian tasks. Tuddenham (1970) has developed several quantitative Piagetian tasks that require some spatial facility and that show sex differences favoring males. The tasks are Perspectives, Water Level, Tracks, Geometric Forms, and House Plans. On the Perspective task, examinees are required to select from among several photographs the one that shows how a small farm would look from various vantage points. The Water Level task involves problems that deal with the principle that water remains gravitationally horizontal regardless of the

tilt of the water's container. The Geometric Forms task involves the identification of flat patterns that can be folded to produce simple three-dimensional forms. Tracks, which involves the least amount of spatial skill, requires the examinee to correctly place a small car painted red on one side and blue on the other at various places on a spiral track. The House Plans task requires the construction of block buildings. Males' mean scores on all tasks except Tracks were higher than females' mean scores (Tuddenham, 1970).

The Water Level task, developed initially by Piaget and Inhelder (1956), was used by Thomas, Jamison, and Hummel, (1973) who found that 31% of adult females but 84% of males had mastered the principle that water remains gravitationally horizontal regardless of the tilt of the water's container. Early studies by Piaget and Inhelder (1956) determined that mastery of this principle occurred by about 12 years of age and that girls lag behind boys at various age levels (Thomas et al., 1973; Liben, Note 7; Thomas, Note 8). And Harris (1978) has suggested that the female lag in this principle's attainment is probably due to the spatial element of the task. Harris' hypothesis has been tested and supported by Geiringer and Hyde (1976). They found correlations between average errors on the Water Level task and performance on a test of spatial orientation of −.83 for 12th-grade males and −.97 for 12th-grade females. Corresponding correlations for 5th-grade males and females were somewhat lower (−.65 for males and −.42 for females) and, although a significant sex difference was found in performance on both tests among 12th graders, none was found among 5th graders. Analysis of covariance revealed that sex differences among 12th graders on the Water Level task disappeared once differences in spatial orientation ability were removed.

Maze tasks. Maze tasks were used as early as 1918 by Porteus in his attempt to design an alternative to measures of general intelligence and verbal ability. Designed for use by human subjects, the Porteus Maze Test (Porteus, 1918) has become an important and widely used testing device within the discipline of psychology (Riddle & Roberts, 1977). A sex difference showing male superior performance on the Porteus Maze Test has been a persistent finding in hundreds of studies since 1918 (Porteus, 1965). As in studies using standard tests of spatial ability, the sex difference does not reliably emerge until after age 11 or 12 (Batalla, 1943; Langhorne, 1948; McNemar, 1942; Porteus, 1965) and has failed to emerge in children younger than 6 years of age (Mattson, 1933; McGinnis, 1929).

Reference is seldom made in discussions of human spatial abilities to the considerable body of evidence that has shown a consistent superiority in maze-learning tasks for male rats (e.g., Barnes et al., 1966; Barrett & Ray, 1970; Cowley & Griesel, 1963; Dawson, 1972; Hubbert, 1915; McNemar & Stone, 1932; Sadownikova-Koltzova, 1926; Tomlin & Stone, 1933; Tryon, 1931).

Summary

We have illustrated the complexity of the problem suggested earlier—that of determining the depth and breadth of the field within which a Spatial factor may be found. Spatial visualization seems to be required in various perceptual-cognitive tasks involving the mental transformation of visual images, and it has been shown to be important for success in college mathematics, especially geometry and algebra.

Spatial orientation enters into such tasks as field dependence – field independence, map reading, and sense of direction. Various Piagetian tasks and maze tasks requiring an aptitude for remaining unconfused by changing orientations of a spatial configuration must certainly involve a strong spatial orientation element. There is an obvious need for further research to clarify the issues presented and to further specify the scope within which the Spatial Visualization and Spatial Orientation factors can be found.

Sources of Variance in Spatial Test Scores

An overwhelming impression conveyed by surveying the spatial abilities literature of the 1960s and 1970s in contrast to the preceding 5 decades is the redirection of interest from factor analytic studies that have con-

clusively distinguished two spatial abilities, visualization and orientation, to both correlational and experimental studies aimed at determining sources of variance in spatial test scores.

One conclusion concerning the factors that account for individual differences in spatial test scores is certain—the empirical evidence is compatible with a relatively broad range of intellectual positions. This evidence is reviewed in four categories, environmental, genetic, hormonal, and neurological, with emphasis on age and sex differences.

Environmental Influences

The importance of experiential factors in the development of spatial skills has been suggested by Berry (1966) in a previously discussed study. Berry compared Eastern Canadian Eskimos from Baffin Island with members of the Temne tribe of Africa on a number of perceptual-cognitive abilities. Eskimos were less field dependent than the Temne and obtained higher mean scores on tests of spatial abilities. Although males in the Temne tribe performed significantly better than Temne females on these measures, there were no significant differences between male and female Eskimos. Unlike Temne females, Eskimo females tend to share equally with males in experiences of hunting. To survive, Eskimo hunters must travel extensively on both land and sea and are required to orient themselves in a relatively featureless array of visual stimuli. Wandering from the home in the activity of hunting presumably fosters a "directional sense" and facilitates spatial skill (Berry, 1966). Although the results seem to suggest the importance of experiential factors in the development of spatial skills, cross-cultural differences are not necessarily environmental, particularly for long isolated groups such as the Temne of Africa and Canadian Eskimos.

Differential training and spatial perception. If experiential factors were important in fostering high spatial abilities, we might expect that training received in this area would result in improved performance on tests of spatial visualization and orientation. There is some evidence that improvement in perceptual judgments occurs as a function of controlled practice and training (Gibson, 1953; Goldstein & Chance, 1965; Kato, 1965; Salkind, 1976; Santos & Murphy, 1960; Van Voorhis, 1941; Witkin, Note 9), although ordinary school curriculum offerings are not always effective in developing spatial perception to the asymptote of an individual's ability (Brinkmann, 1966; Brown, 1954; Mendicino, 1958; Ranucci, 1952).

Blade and Watson (1955) reported significant increases by students on a test of spatial visualization during an engineering course. Positive effects have also been demonstrated by Brinkmann (1966). A small number of eighth-grade boys ($n = 14$) and girls ($n = 13$) were instructed in selected concepts of geometry and mental pattern folding during mathematics classes over a 3-week period. Posttraining test scores on the Space Relations Subtest of the Differential Aptitude Test (Bennett et al., 1974) were significantly higher than pretraining test scores for the experimental groups but not for a control group. Dailey and Neyman (Note 10) attempted to train vocational high school students on items similar to those found on two- and three-dimensional tests of orientation and visualization. Posttraining test scores obtained at the end of the academic year were significantly higher than pretraining test scores. These gains were shown to be greater for the trained group of students than for a control group.

Conflicting evidence in the literature is provided by Faubian, Cleveland, and Hassell (1942) who reported no differences on the Surface Development Test between a group of Air Corps recruits who had received training in drafting and blueprint reading and a matched control group. Churchill, Curtis, Coombs, and Hassell (1942) found slight but insignificant gains on the same test after a 9-week training course in engineering drawing. Myers (Note 11) found that Naval cadets who had received training in mechanical drawing scored no higher on spatial tests than cadets who had received no such training. And Ranucci (1952) and Brown (1954) have independently shown that courses in high school geometry did not result in increased performance on spatial tests.

An increase in mean performance levels after training on spatial tasks, even if it were a consistent finding, does not explain the sex difference. If we assume that the female deficit in human spatial abilities results from differential learning and that males are closer to the asymptote of their ability than females, then females should respond more favorably than males to training. This hypothesis is not strongly supported by available evidence. Unfortunately, sex differences in response to training were not systematically examined in the studies reviewed previously. However, a few studies are available. Preliminary results from Drauden's University of Minnesota dissertation research show no such affect (Drauden, Note 12). Teegarden (1942) examined the effect of increased time limits on test scores and found that increasing the time allowed to complete a form-board test did not significantly effect female performance. McGee (1978a) found no evidence for a differential response to training and practice on the Mental Rotations Test by females in comparison with males. Smith (1948) also found that females did not differentially respond to training. Female spatial test performance was not raised to the level of male performance as the result of training received in technical school. Thomas et al. (1973) reported sex differences for mastery of the principle that the surface of still water remains horizontal regardless of the tilt of the water's container. Whereas 84% of the males ($n = 62$) were aware of the principle, only 31% of the females ($n = 91$) were similarly aware. Training was successful in teaching the concept of "horizontality" to only 12 of the 63 females who were initially unaware. And McGee (Note 13) found that first-grade boys benefited more than girls from training on the Piagetian Water Level task.

Empirical research does not strongly support the hypothesis of a differential response to training by females than males. However, as pointed out by Sherman (1967), because of the unknowns involved in assuming what is relevant activity in increasing spatial abilities, it is difficult to know whether the sexes do in fact receive differential practice. Many activities in our culture are sex typed.

> Very few girls are found in the high school classes of mechanical drawing, analytical geometry, and shop. Spare-time activities of tinkering with the car, sports, model building, driving a car, direction finding, and map reading are sex-typed and might also be sources of differential practice. (Sherman, 1967, p. 295)

It seems highly likely that these and similar activities are involved in fostering spatial skills. To uncover this effect experimentally will require research designs more sensitive than those that have so far been used.

Genetic Influences

Accumulated evidence shows that spatial abilities are equally as or more heritable than verbal ability (Blewett, 1954; Block, 1968; Bock, 1973; DeFries et al., 1974, 1976; McGee, 1978c; Osborne & Gregor, 1968; Park et al., 1978; Vandenberg, 1962, 1967, 1968, 1969, 1971; Vandenberg, Stafford, & Brown, 1968; Williams, 1975; Bramble, Bock, & Vandenberg, Note 14; Thurstone, Thurstone, & Strandskov, Note 15) and much less correlated with traditional measures of environmental quality such as level of education and SES (Bock & Vandenberg, 1968; Marjoribanks, 1972; McGee, 1977; Vandenberg, 1971). A number of studies have suggested that spatial abilities may be enhanced by an X-linked, recessive gene (Bock & Kolakowski, 1973; Goodenough et al., 1977; Guttman, 1974; Hartlage, 1970; Stafford, 1961; Yen, 1975) and this hypothesis has served as a tentative explanation for the mode of genetic transmission and the sex difference in spatial test performance.

Spatial abilities and X-linked inheritance. Traits effected by the transmission of a single gene on the X chromosome are said to be X-linked and are determined to be either dominant or recessive based on the relative frequency of effected males and females in the population. If the X-linked trait were recessive, more males than females would be affected, whereas if the X-linked trait were dominant, more affected females than males would be expected. This is true because in a population at equilibrium, one

third of the X-linked genes are carried by males and two thirds are carried by females, since females inherit two X chromosomes (one from each parent) and males inherit only one. A recessive, X-linked trait will be expressed in hemizygous recessive males and homozygous recessive females but not in hemizygous dominant males nor in heterozygous or homozygous dominant females. Females with a double recessive genotype would be expected to occur in the population with a frequency q^2—the square of the frequency of males carrying the single recessive allele.

Where the frequency of the recessive, spatial-enhancing allele q equals either 0 or 1.00, the absolute sex difference $(q - q^2)$ will be 0 (Jensen, 1975). As the value q departs from 0 or 1.00, the absolute sex difference will increase. A gene frequency of .5 of the spatial-enhancing allele maximizes the sex difference, which sets the ratio of enhanced females to males at 1:2. Thus O'Connor's (1943) observation that only one fourth of all females score above the male median on tests of spatial ability, a finding replicated by numerous investigators (e.g., Bock & Kolakowski, 1973; Bouchard & McGee, 1977; Loehlin, Sharan, & Jacoby, 1978; Yen, 1975), is in accordance with the X-linked, recessive model.

In addition to fulfilling the basic requirement of explaining the greater proportion of spatializing males than females, the genetic X-linkage model predicts a characteristic pattern of family correlations different from that expected for an autosomal, polygenic trait. The model predicts a higher father–daughter than father–son correlation and a higher mother–son than mother–daughter correlation. Opposite sex pairs of siblings will tend to show less similarity than pairs of sisters, whereas the correlation between pairs of brothers should be an intermediate value. Sisters will be most similar because for brothers, the one X chromosome may be either of the mother's two, whereas for sisters the paternal X chromosome is identical (Hogben, 1932; Mather & Jinks, 1963; McKusick, 1964). Thus, under random mating and a recessive gene frequency of .5 (the frequency that best explains the mean sex difference and the shape of the male and female distribution), the expected order of family correlations is as follows: r sister–sister $>$ r mother–son $=$ r father–daughter $>$ r brother–brother $>$ r mother–daughter $>$ r brother–sister $>$ r father–son. The theoretically expected correlation, based on X-linkage, of .00 between fathers and sons is not easily predicted from environmental hypotheses in which modeling effects and shared experiences would ordinarily be expected to lead to higher same-sex than opposite-sex parent–child correlations. Consequently, the original studies supporting the X-linkage hypothesis (Bock & Kolakowski, 1973; Hartlage, 1970; Stafford, 1961) generated considerable interest, despite the fact that they were based on rather small samples and failed to provide correlations among siblings.

Several recent tests of the hypothesis have been conducted. This research is summarized in Table 2, which shows age-corrected family correlations for tests of spatial visualization and spatial orientation abilities, along with results on two miscellaneous tests that involve a spatial component but that are not ordinarily recognized as measures of visualization or orientation. Several of the studies used numerous tests, and some of the tests were used in more than one study. Thus, in addition to comparing similarities (and dissimilarities) among tests, it is possible to examine the extent to which the same tests yield similar correlation patterns across samples.

Only two studies in the literature (Bouchard & McGee, 1977; Loehlin et al., 1978) report the complete array of family correlations consisting of both parent–child and sibling correlations. In each of these studies, the father–son correlation equals or exceeds the mother–son correlation; therefore these results do not conform to the expected pattern. In Bouchard and McGee's (1977) study the difference between the brother–brother and the sister–sister correlations is highly significant ($p < .005$) in the direction opposite that predicted by the X-linkage model. The expected pattern of parent–child correlations (with opposite-sex parent–child correlations highest, the mother–daughter correlation intermediate, and the father–son correlation lowest), although obtained in the

original studies supporting the X-linkage hypothesis (Bock & Kolakowski, 1973; Hartlage, 1970; Stafford, 1961), has not been obtained for any spatial tests used in four recent, larger family studies (Bouchard & McGee, 1977; DeFries et al., 1976; Loehlin et al., 1978; Park et al., 1978). These results provide weak evidence, if any, for the spatial-enhancing effect of the X-linked recessive gene postulated by Stafford (1961). The idea, however, that different assessments of spatial abilities (e.g., two-dimensional vs. three-dimensional tasks, rotation vs. transformation of spatial objects, analytic vs. gestalt processing, visualization vs. orientation) have different genetic structures should appeal to those investigators who are reluctant to abandon the search for the spatial gene. An important contribution to this literature would be a family study providing the full array of intrafamilial correlations for spatial visualization and spatial orientation tests, including the three tests that have shown the pattern of correlations predicted by the X-linked model: the Identical Blocks Test (Stafford, 1961), the Differential Aptitude Space Test (Hartlage, 1970), and the adapted Guilford-Zimmerman Spatial Visualization Test (Bock & Kolakowski, 1973). It is not at all clear, from the few attempts to address these issues (Guttman, 1974; Loehlin et al., 1978; Yen, 1975) just what kinds of spatial test performance the X-linked gene should be expected to influence most. "Clearly, the final word is not yet in" (Loehlin et al., 1978, p. 40).

Hormonal Influences

To what extent are hormonal differences in males and females responsible for the observed sex difference in spatial test performance? A small body of studies has addressed this question. Petersen (1976), for example, demonstrated a curvilinear relationship between physical androgenicity and certain aspects of cognitive functioning, namely, verbal fluency and spatial ability. Measures in both boys and girls at ages 13, 16, and 18 were made of sex hormone influence inferred from degree of secondary sex characteristic development, which is known to be under gonadal hormone influence (Tanner, 1969).

Scores were available from two measures of spatial ability (including the Primary Mental Abilities Space Test, Thurstone & Thurstone, 1965) and from two measures of verbal fluency. Her results showed that less androgenized (less masculine) males (assessed on the basis of various physical characteristics including hip size, shoulder width, and muscle strength) scored higher on the spatial tests than boys with a more androgenized (more masculine) body type. The positive correlations were highest for the 18 year olds, lowest for the 13 years olds, and intermediate for the 16 year olds. These findings support those presented by Broverman, Klaiber, Kobayashi, and Vogel (1968), who found that more androgenized males scored higher on verbal fluency than on spatial ability tests. For females, however, Broverman et al. found different results: Females with high androgen levels (again rather crudely determined on the basis of physical characteristics such as narrow hips, wide shoulders, solid muscles, and small breasts) had higher spatial scores than females with low androgen levels.

Other evidence has accumulated that highly masculinized males have a tendency toward lower spatial scores. For example, Klaiber, Broverman, and Kobayashi (1967) tested male college students using two simple repetition tasks and two spatial tasks. Scores on these spatial tasks were correlated with measures of masculinity (large chest and biceps and pubic hair distribution). Results showed a positive correlation with performance on the simple tasks of repetition but a negative correlation with performance on the spatial tasks. In another study, Ferguson and Maccoby (cited in Maccoby, 1966) found that boys with high spatial scores were rated by their peers as less masculine than boys with low scores.

How are these data to be interpreted? High body androgenization is associated with *low* spatial scores among males (Ferguson & Maccoby, cited in Maccoby, 1966; Klaiber et al., 1967; Petersen, 1976) and with ***high*** spatial scores among females (Broverman et al., 1968; Petersen, 1976). It might be

Table 2
Summary of Age-Corrected Family Correlations for Tests of Spatial Visualization and Spatial Orientation Abilities

Test	Study	No. of families	Parent and parent–child correlations					Sibling correlations		
			F–M	F–S	M–D	M–S	F–D	B–Ss	B–B	Ss–Ss
		Spatial visualization tests								
Identical Blocks Test[a]	Stafford (1961)	104	.03	.02	.14	.31	.31	—	—	—
DAT Space Test[b]	Hartlage (1970)	25	—	.18	.25	.39	.34	—	—	—
Guilford-Zimmerman Spatial Visualization[c]	Bock and Kolakowski (1973)	167	.26	.15	.12	.20	.25	—	—	—
Form Board Test (Form CC)[d]	Yen (1975)	—	—	—	—	—	—	.29	.51	.68
Paper-Folding Test[d]	Yen (1975)	—	—	—	—	—	—	.25	.35	.48
Paper-Folding Test[d]	Loehlin, Sharan, and Jacoby (1978)	192	.09	.27	.21	.24	.30	.17	.44	.44
Paper Form Board Test[e]	DeFries et al. (1976)[f]									
	AEA sample	739	.20	.27	.36	.30	.40	—	—	—
	AJA sample	244	.06	.24	.24	.26	.21	—	—	—
Paper Form Board Test[e]	Park et al. (1978)	209	—	.59	.57	.63	.53	—	—	—
Mental Rotation Test[g]	Yen (1975)	—	—	—	—	—	—	.27	.32	.41
Mental Rotation Test[g]	DeFries et al. (1976)[f]									
	AEA sample	739	.10	.15	.32	.16	.23	—	—	—
	AJA sample	244	.02	.30	.10	.14	.34	—	—	—
Mental Rotation Test[g]	Park et al. (1978)	209	—	.22	.46	.26	.41	—	—	—
Mental Rotation Test[g]	Bouchard and McGee (1977)	200	.06	.23	.16	.20	.17	.33	.50	.21
		Spatial orientation tests								
PMA Spatial Relations[a]	Yen (1975)	—	—	—	—	—	—	.30	.07	.46
Cube Comparisons Test[d]	Loehlin, Sharan, and Jacoby (1978)	192	.01	.16	.19	.04	.17	.32	.43	.14
Card Rotations Test[d]	DeFries et al. (1976)[f]									
	AEA sample	739	.13	.26	.36	.19	.22	—	—	—
	AJA sample	244	.04	.26	.09	.12	.11	—	—	—
Card Rotations Test[d]	Park et al. (1978)	209	—	.12	.61	.36	.54	—	—	—
Card Rotations Test[d]	Loehlin, Sharan, and Jacoby (1978)	192	.28	.27	.40	.27	.32	.24	.44	.52

Table 2 (*continued*)

Test	Study	No. of familes	Parent and parent–child correlations					Sibling correlations		
			F–M	F–S	M–D	M–S	F–D	B–Ss	B–B	Ss–Ss
Miscellaneous spatial tests										
Raven's Progressive Matrices	Guttman (1974)	100	.26	.36	.39	.24	.23	—	—	—
Raven's Progressive Matrices	DeFries et al. (1976)[f]									
	AEA sample	739	.22	.20	.29	.38	.27	—	—	—
	AJA sample	244	.20	.10	.17	.22	.26	—	—	—
Raven's Progressive Matrices	Park et al. (1978)	209	—	.33	.51	.25	.39	—	—	—
Hidden Patterns Test[a]	DeFries et al. (1976)[f]									
	AEA sample	739	.23	.21	.32	.20	.27	—	—	—
	AJA sample	244	−.05	.10	.18	.12	.26	—	—	—
Hidden Patterns Test[a]	Park et al. (1978)	209	—	.51	.65	.58	.56	—	—	—
Hidden Patterns Test[a]	Loehlin, Sharan, and Jacoby (1978)	192	.21	.40	.22	.44	.38	.39	.76	.55
Hidden Patterns Test[a]	McGee (1978c)	200	.26	.07	.35	.23	.39	.11	−.03	.20
Expected correlations for X-linked inheritance ($q = .5$)[h]			.00	.00	.38	.58	.58	.13	.50	.67

Note. F = father; M = mother; S = son; D = daughter; B = brother; Ss = sister; DAT = Differential Aptitude Test; PMA = Primary Mental Abilities.

[a] Thurstone and Thurstone (1941).

[b] Bennett, Seashore, and Wesman (1974).

[c] Guilford and Zimmerman (1953).

[d] French, Ekstrom, and Price (Note 2) and Ekstrom, French, and Harman (Note 3).

[e] Likert and Quasha (1970).

[f] Father—mother correlations, from Johnson (1976), are based on 555 American couples of European ancestry (AEA) and 148 American couples of Japanese ancestry (AJA).

[g] Vandenberg (1975) and Vandenberg and Kuse (1978).

[h] Assuming random mating and no nonheritable variation.

that spatial abilities are facilitated not by any absolute level of androgen but rather by an optimal estrogen–androgen balance. Unfortunately there exists a paucity of evidence to support this contention. One test comes from a study of Kwashiorkor feminized males (Dawson, 1967a, 1967b). West African children suffering from Kwashiorkor (a disease resulting from prolonged subsistence on diets deficient in protein) and the accompanying protein-deficiency-induced endocrinological dysfunction of the liver, which prevents the normal inactivation of the production of estrogen in the male (Stuart-Mason, 1963), exhibited lower spatial and numerical ability and greater field dependence, as compared to normal controls, in addition to demonstrating greater verbal ability. And the curvilinear nature of the relationship between body androgenicity and spatial ability found by Petersen (1976) indicates that at least a minimum androgen level is required for normal spatial ability. It follows from these findings that the superior spatializer of either sex is less sexually differentiated than are nonspatializers. That is, the estrogen–androgen balance would be optimal and consequently spatial abilities would be highest for males low in androgen and for females high in androgen. We might even suggest that for females, the more androgen one has the better, thus explaining why individuals with Turner's syndrome (phenotypic females, the majority of whom has the single X chromosomes XO rather than the normal female XX pairs and no gonadal hormones) demonstrate poorer spatial abilities (Alexander, Ehrhardt, & Money, 1966; Alexander & Money, 1966; Alexander, Walker, & Money, 1964; Garron, 1970, 1977; Money, 1963; Money & Granoff, 1965; Serra, Pizzamiglio, Boari, & Spera, 1978; Silbert, Wolff, & Lilienthal, 1977), poorer direction sense (Alexander et al., 1964), and greater field dependence (Serra et al., 1978) than both males who also have single X chromosomes and genetically normal females.

The questions raised in this section point to the obvious need for research aimed at clarifying the relationship between somatic androgenization and the evidence reviewed earlier for an X-linked recessive gene influence on spatial skill enhancement. The finding of better spatial skill in late-maturing, less androgenized boys than in early-maturing, more androgenized boys (Broverman et al., 1968) suggests the operation of an X-linked gene controlling the timing of release of androgen rather than the expression of spatial skill directly (Bock & Kolakowski, 1973).

The suggestion made is that there exists a sex-linked influence on within-sex variation in somatic androgenicity. However, this is a question that awaits empirical investigation. Future research might be aimed at determining the extent of this influence on between-sexes variation in spatial abilities. The methods of measuring body androgenization used by Petersen (1976) and by Broverman et al. (1968), based on the analysis of physical characteristics including muscle development, body shade, genital or breast size, and pubic hair distribution obtained by rating photographs, are crude and imprecise indices of sexual differentiation controlled by estrogen–androgen balance. Until more direct methods of hormonal assay are employed on larger samples, the precise nature of the relationship between spatial abilities and hormonal balance will remain an open question.

Neurological Influences

Neurological studies showing variations in the lateral organization of the human brain provide experimental evidence for a structural source of the variation in human spatial abilities. Recent work on hemispheric specialization suggests (a) that the right cerebral hemisphere is specialized for spatial processing and (b) that males have greater hemisphere specialization than females. These conclusions are supported from several different types of evidence to be reviewed.

Hemispheric specialization. Language function was the first higher mental process found to be asymmetrically represented in the human brain, and it remains the best documented case of hemispheric specialization (Nebes, 1974). Recent studies have established anatomical bases for the special-

ization of both verbal and nonverbal information processing in human subjects. Kimura (1961) was probably the first investigator to employ Broadbent's (1954) technique of dichotic listening for the examination of hemispheric specialization when she demonstrated that when pairs of contrasting digits were presented simultaneously to the right and left ears, those presented to the right ear were more accurately reported. Right ear advantage for the processing of easy-to-verbalize stimuli (e.g., numbers, words, and letters) has since been confirmed (Milner, Taylor, & Sperry, 1968; Sparks & Geschwind, 1968; Studdert-Kennedy & Shankweiler, 1970).

Conversely, a left ear (right hemisphere) advantage for the processing of difficult-to-verbalize stimuli (e.g., melodies, sonar signals, and abstract patterns of sound) has also been demonstrated (Curry, 1967; Kimura, 1964, 1966; Shankweiler, 1966; Spreen, Benton, & Fincham, 1965; Vignolo, 1969; Chaney & Webster, Note 16).

Other data demonstrate convincingly that each cerebral hemisphere primarily subserves its contralateral limb and binocular visual hemifield (Buffery & Gray, 1972) and that in about 96% of the normal adult population, cerebral dominance for verbal functions (i.e., tasks requiring semantic memory, manipulation, and production) is subserved by the left hemisphere, whereas the right hemisphere predominates in subserving nonverbal functions (i.e., tasks requiring perception and manipulation of visual images) (Bogen & Gazzaniga, 1965; Buffery, 1968; Buffery & Gray, 1972; Levy, 1976a; Searleman, 1977; Witelson, 1976).

Several investigations of patients with unilateral brain lesions have demonstrated spatial abilities to be more affected by right than by left cerebral injury (Kimura, 1967; McFie, Piercy, & Zangwill, 1950; Milner, 1962). Costa and Vaughan (Note 17) found that right-lesion patients ($n = 18$) scored significantly lower ($p < .05$) than left-lesion patients ($n = 18$) on the Block Design subtest of the Wechsler Adult Intelligence Scale, with both normal and extended time limits. Similarly, it has been demonstrated that patients who have suffered the loss of their left temporal lobe show impaired memory for verbal materials but nonsignificant performance decrements on tasks such as memory for faces (Milner et al., 1968) and maze learning (Corkin, 1965; Milner, 1965). In a more recent study, Kershner and King (1974) examined laterality of cognitive functions among hemiplegic children and found similar results. Twenty-one children were administered the Wechsler Intelligence Scale for Children (WISC) and the Reitan-Indiana Neurological Test. The left hemiplegics (right-brain-damaged) children ($n = 7$) were poorer than normals on visuoperceptual performance tasks ($p < .05$) but showed no significant impairment, relative to normals, on any of the WISC verbal tests. Right hemiplegics (left-brain-damaged) children ($n = 7$) were poorer than normals in verbal intelligence ($p < .05$).

Although sample sizes are small among clinical populations, clinical studies of the effects of unilateral brain damage, reviewed previously, and commissurotomy (Sperry, 1968, 1973; Sperry, Gazzaniga, & Bogen, 1969) on verbal and spatial tasks corroborate tachistoscopic perceptual data and provide direct evidence for the conclusion that the right cerebral hemisphere is specialized for spatial processing.

Hemispheric specialization and sex differences in spatial abilities. To what extent do sex differences in hemispheric specialization underly male superiority on tasks requiring spatial abilities? A review of clinical and experimental data indicates that the right cerebral hemisphere is specialized for spatial processing and that the cerebral hemispheres of males and females tend to show differences in specialization for verbal and spatial functions. The conclusions that males have greater right hemisphere specialization than females is supported by data from tachistoscopic perceptual studies (Ehrlichman, 1972; Kimura, 1969, 1973; McGlone & Davidson, 1973), clinical studies (Lansdell, 1962, 1968a, 1968b, 1973; McGlone & Kertesz, 1973), and studies of anatomical differences between the sexes (Geschwind, 1974; Lansdell & Davie, 1972; Wada, 1974; Witelson & Pallie, 1973). Much of this literature has been reviewed elsewhere (cf. Buffery & Gray,

1972; Harris, 1978; Harshman & Remington, 1976).

The major opposition to the conclusion that males have greater right hemisphere specialization and thus greater spatial ability than females has been proposed by Buffery and Gray (1972). Their theory that females are more lateralized than males for both language and spatial skills is supported mainly from developmental data on children. However, as Buffery and Gray themselves point out, "sex differences in children are difficult to interpret when there is an advantage in favor of girls, since this may always be due to their general maturational advantage over boys" (p. 131). Moreover, recent developmental studies of hemispheric specialization are consistent with the conclusion that boys have greater right hemisphere specialization than girls and that girls are more bilateral in their cerebral representation of verbal and spatial functions.

Knox and Kimura (1970), for example, studied dichotic listening to nonverbal stimuli (environmental and animal noises such as dish washing, phone dialing, clock ticking, dog barking). Subjects were 80 right-handed children between 5 and 8 years of age. Males showed a greater left ear (right hemisphere) superiority than females across all ages.

Witelson (1976) presented children ranging in age from 3 to 13 years with a tactual version of the dichotic recognition technique. All children were originally assessed as being right-handed. Examinees were instructed to touch unfamiliar four- to eight-sided shapes and then to identify the forms by pointing to a visual display of a group of shapes. Boys ($n = 165$) at age 5 and beyond showed a significant left-hand (right hemisphere) advantage; there was no hand difference for the 3- to 4-year-olds. Girls ($n = 165$) showed significant left-hand superiority but not until after age 13. Witelson concluded that the right hemisphere may be specialized for spatial processing (the detection of shapes) earlier in boys than in girls. Other evidence of earlier right hemisphere specialization in boys than girls for processing of nonauditory, tactual configurations has been provided from the investigation of Braille reading in both normal (Rudel, Denckla, & Spalten, 1973) and blind subjects (Hermelin & O'Connor, 1971a, 1971b). Since Braille characters are symbols of alphabet letters, a left hemisphere (right-hand advantage for reading Braille might be expected. It has been observed, however (Hermelin & O'Connor, 1971b), that right-handed blind individuals have a clear advantage for reading Braille with the fingers of their left hand. Hermelin and O'Connor (1971a) have suggested that for the blind, Braille symbols are processed as spatial configurations, not as linguistic symbols, and are as a result processed by the right cerebral hemisphere. A direct test of this hypothesis has been provided by Rudel et al. who examined normal children. Right-handed males and females ($n = 80$) between 7 and 14 years of age learned 12 Braille letters, 6 with each hand. Results indicated that 7- and 8-year-old boys performed equally well with both hands in the Braille, paired-associate learning task but that the girls of the same age showed right-hand superiority. Left-hand (right hemisphere) superiority emerged for both boys and girls at later ages (13 and 14 years), but the difference between right- and left-hand scores was statistically significant only for the boys, indicating an earlier and perhaps superior pattern of right hemisphere development in boys than girls.

Another opposing hypothesis to that of greater right hemisphere specialization in males than females has been proposed by Harris (1978). According to Harris, "the male eventually equals and then surpasses the female in degree of left hemisphere lateralization, so that in adulthood, language in females is bilaterally represented," thus impeding her spatial ability (p. 460). Support for Harris' first postulate—that of greater bilateralization of language function in females than males—is provided from studies of normal (Bryden, 1966; Remington, Krashen, & Harshman, Note 18) as well as clinical populations (Lansdell, 1961, 1962; McGlone & Kertesz, 1973). Harris dismisses developmental studies of hemispheric specialization, however, which consistently show earlier and greater left hemisphere specialization of language function in

girls than boys (Buffery, 1970, 1971a, 1971b, 1971c; Kimura, 1963; Pizzamiglio & Cecchini, 1971; Bryden, Allard, & Scarpino, Note 19).

Harris (1978) provides support for his second postulate—that bilateral cerebral representation impedes spatial skills—mainly on the basis of studies of left-handers. The assumption is that left-handers tend to be less well lateralized (more bilateral) than right-handers in cerebral representation of verbal and spatial functions (Bryden, 1966; Goodglass & Quadfasel, 1954; Remington et al., Note 18). The implication is that left-handers, like females, should score lower on tests of spatial ability than right-handers, since they are less well lateralized. Studies supporting the relationship between left-handedness and deficits on spatial tasks (James, Mefferd, & Wieland, 1967; Levy, 1969, 1976b; McGlone & Davidson, 1973; Miller, 1971; Nebes, 1971; Nebes & Briggs, 1974; A. Silverman, Adevai, & McGough, 1966) are based on small samples, and differences associated with sex are not always examined.

Numerous other studies (Annett & Turner, 1974; Fagan-Dubin, 1974; Kutas, McCarthy, & Donchin, 1975; McGee, 1976, 1978b; Newcombe & Ratcliff, 1973; Sherman, in press) report conflicting evidence regarding the prediction of poorer overall performance on spatial tasks by left- than right-handers. As noted by Hardyck and Petrinovich (1977), the data indicating that left-handedness is associated with deficits of various kinds is far from compelling.

A related problem associated with Harris' (1978) hypothesis is the assumption that left-handers are more bilateral than right-handers in their cerebral representation of verbal and spatial functions. Bilaterality of cerebral function seems to be present in the left-handed only when there is a family history of left-handedness (Hardyck & Petrinovich, 1977) and is *not* a characteristic of the nonfamilial left-handed individuals who as a group tend to be organized for cerebral specialization exactly as are the right-handed.

In summary, the clinical and experimental neurological literature suggests conclusively that the right cerebral hemisphere is specialized for spatial processing and that males have greater right hemisphere specialization than females. Further research is needed to determine the causal relationship, if any, between sex differences in hemisphere specialization and sex differences in spatial abilities.

Conclusion

Six conclusions are warranted. First, a plethora of factor analytic studies since the 1930s have provided strong and consistent support for the existence of at least two distinct spatial abilities—visualization and orientation. Spatial visualization is the ability to mentally rotate, manipulate, and twist two- and three-dimensional stimulus objects. Spatial orientation ability includes the comprehension of the arrangement of elements within a visual stimulus pattern, the aptitude to remain unconfused by the changing orientations in which a spatial configuration may be presented, and an ability to determine spatial orientation with respect to one's body. Second, visualization and orientation abilities are more highly correlated with success in a number of technical, vocational, and occupational domains than is verbal ability, making them important variables in applied psychology. Third, sex differences in various aspects of perceptual-cognitive functioning (e.g., mathematics and field independence) are a secondary consequence of differences with respect to spatial visualization and spatial orientation abilities. Fourth, sex differences on tests of spatial visualization and orientation as well as on numerous tasks requiring these abilities do not reliably appear until puberty. Fifth, spatial abilities are known to be influenced almost as much by genetic factors as is verbal ability in all populations studied; however, the X-linked recessive gene hypothesis that has served as a tentative explanation for sex differences in spatial abilities and for the mode of genetic transmission is not supported strongly in recent studies. Sixth, the development of sex differences in spatial skills is likely related to sex differences in the development of hemisphere specialization. Recent work in hemisphere specialization demonstrates conclusively that the right cerebral

hemisphere is specialized for spatial processing and that males have greater right hemisphere specialization than females.

Reference Notes

1. Thurstone, L. L. *Some primary abilities in visual thinking* (Report No. 59). Chicago: University of Chicago, Psychometric Laboratory, 1950.
2. French, J. W., Ekstrom, R. B., & Price, L. A. *Kit of reference tests for cognitive factors.* Princeton, N.J.: Educational Testing Service, 1963.
3. Ekstrom, R. B., French, J. W., & Harman, H. H. *Manual for kit of factor referenced cognitive tests.* Princeton, N.J.: Educational Testing Service, 1976.
4. Myers, C. T. *Some observations of problem solving in spatial relations tests* (ETSRB58-16). Princeton, N.J.: Educational Testing Service, 1958.
5. Lane, J. B. *Imagery and behavior change.* Unpublished manuscript, University of Minnesota, 1974.
6. Canavan, D. Field dependence in children as a function of grade, sex, and ethnic group membership. In Harold B. Gerard (Chair), *School desegregation and achievement-related attitudes.* Symposium presented at the 77th annual convention of the American Psychological Association, Washington, D.C., August 1969.
7. Liben, L. S. *Operative understanding of horizontality and its relation to long-term memory.* Paper presented at the meeting of the Society for Research in Child Development, Philadelphia, 1973.
8. Thomas, H. *The development of water-level representation.* Paper presented at the meeting of the Society for Research in Child Developmen, Minneapolis, 1971.
9. Witkin, H. A. *The effect of training and of structural aids on performance in three tests of spatial orientation* (Report No. 80). Washington, D.C.: Civil Aeronautics Association, Division of Research, 1948.
10. Dailey, J. T., & Neyman, C. A. *Development of a curriculum and materials for teaching basic vocational talents.* (Final Report for Office of Education, Contrast No. OE-5-85-023). Washington, D.C.: George Washington University, Education Research Project, 1967.
11. Myers, C. T. *The effects of training in mechanical drawing on spatial relations test scores as predictors of engineering drawing grades* (ETS RM 58-4). Princeton, N.J.: Educational Testing Service, 1958.
12. Drauden, G. *Training effects on sex differences in spatial abilities.* Unpublished manuscript, University of Minnesota, 1978.
13. McGee, M. G. *Female superiority before pubert[y] on the Piagetian water level task.* Manuscript i[n] preparation, 1979.
14. Bramble, W. J., Bock, R. D., & Vandenber[g], S. G. *Components of heritable variation in th[e] Primary Mental Abilities Tests* (Research Rep.) Boulder: University of Colorado, Institute o[f] Behavioral Genetics, 1970.
15. Thurstone, T. G., Thurstone, C. C., & Strand-skov, H. H. *A psychological study of twin[s]* (Report No. 4). Durham, N.C.: University o[f] North Carolina, Psychometric Laboratory, 195[?]
16. Chaney, R. B., & Webster, J. C. *Informatio[n] in certain multidimensional signals* (Report No. 1339). San Diego, Calif.: U.S. Navy Electro[nics] Lab, 1965.
17. Costa, L. D., & Vaughan, H. *Verbal and perceptual performance in patients with cerebr[al] lesions.* Paper presented at the 68th annual convention of the American Psychological Association, Chicago, August 1960.
18. Remington, R., Krashen, S., & Harshman, R. A. *A possible sex difference in degree of lateralization of dichotic stimuli.* Paper presented at the meeting of the Acoustical Society of America, Los Angeles, 1973.
19. Bryden, M. P., Allard, F., & Scarpino, F. *The development of language lateralization and speech perception.* Unpublished manuscript, University of Waterloo, Ontario, Canada, 1973.

References

Alexander, D., Ehrhardt, A. A., & Money, J. Defective figure drawing, geometric and human, in Turner's syndrome. *Journal of Nervous and Mental Disease,* 1966, *142,* 161–167.

Alexander, D., & Money, J. Turner's syndrome and Gerstmann's syndrome: Neuropsychologic comparisons. *Neuropsychologia,* 1966, *4,* 265–273.

Alexander, D., Walker, H. T., Jr., & Money, J. Studies in direction sense: I. Turner's syndrome. *Archives of General Psychiatry,* 1964, *10,* 337–339.

Alexander, W. P. Intelligence, concrete and abstract. *British Journal of Psychology Monograph Supplement,* 1935, No. 29.

Anastasi, A. *Differential psychology: Individual and group differences in behavior* (3rd ed.). New York: Macmillan, 1958.

Andrew, D. M. An analysis of the Minnesota Vocational Test for clerical workers: I. *Journal of Applied Psychology,* 1937, *21,* 18–47. (a)

Andrew, D. M. An analysis of the Minnesota Vocational Test for clerical workers: II. *Journal of Applied Psychology,* 1937, *21,* 139–172. (b)

Annett, M., & Turner, A. Laterality and the growth of intellectual abilities. *British Journal of Educational Psychology,* 1974, *44,* 37–44.

Backman, M. E. Patterns of mental abilities of adolescent males and females from different ethnic

and socioeconomic backgrounds. *Proceedings of the 79th Annual Convention of the American Psychological Association,* 1971, *6,* 511–512. (Summary)

Backman, M. E. Patterns of mental abilities: Ethnic, socioeconomic, and sex differences. *American Educational Research Journal,* 1972, *9,* 1–12.

Barnes, R. H., et al. Influence of nutritional deprivations in early life on learning behavior of rats measured by performance in a water maze. *Journal of Nutrition,* 1966, *89,* 399–410.

Barrett, R. J., & Ray, O. S. Behavior in the open field, Lashley III maze, shuttle box, and Sidman avoidance as a function of strain, sex, and age. *Developmental Psychology,* 1970, *3,* 73–77.

Batalla, M. The maze behavior of children as an example of summative learning. *Journal of Genetic Psychology,* 1943, *63,* 199–211.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. *Manual for the differential aptitude tests: Forms S and T* (5th ed.). New York: The Psychological Corporation, 1974.

Berry, J. W. Temne and Eskimo perceptual skills. International Journal of Psychology, 1966, *1,* 207–229.

Betts, G. H. The distribution and functions of mental imagery. *Contributions to Education Series* (Whole No. 26). New York: Columbia University Press, Teachers College, 1909.

Bieri, J., Bradburn, W., & Galinsky, M. Sex differences in perceptual behavior. *Journal of Personality,* 1958, *26,* 1–12.

Bigelow, G. Field dependence – field independence in 5 to 10 year old children. *Journal of Educational Research,* 1971, *64,* 397–400.

Blade, M. F., & Watson, W. S. Increase in spatial visualization test scores during engineering study. *Psychological Monographs,* 1955, *69*(1, Whole No. 397).

Blewett, D. B. An experimental study of the inheritance of intelligence. *Journal of Mental Science,* 1954, *100,* 922–933.

Block, J. B. Hereditary components in the performance of twins on the WAIS. In S. G. Vandenberg (Ed.), *Progress in human behavior genetics.* Baltimore, Md.: Johns Hopkins University Press, 1968.

Bock, R. D. Word and image: Sources of the verbal and spatial factors in mental test scores. *Psychometrika,* 1973, *38,* 437–457.

Bock, R. D., & Kolakowski, D. Further evidence of sex-linked major-gene influence on human spatial ability. *American Journal of Human Genetics,* 1973, *25,* 1–14.

Bock, R. D., & Vandenberg, S. G. Components of heritable variation in mental test scores. In S. G. Vandenberg (Ed.), *Progress in human behavior genetics.* Baltimore, Md.: John Hopkins University Press, 1968.

Bogen, J. E., & Gazzaniga, M. S. Cerebral commisurotomy in man: Minor hemisphere dominance for certain visio-spatial functions. *Journal of Neurosurgery,* 1965, *23,* 394–399.

Bogo, N., Winget, C., & Gleser, G. C. Ego defenses and perceptual styles. *Perceptual and Motor Skills,* 1970, *30,* 599–604.

Borich, G. D., & Bauman, P. M. Convergent and discriminant validation of the French and Guilford-Zimmerman spatial orientation and spatial visualization factors. *Educational and Psychological Measurement,* 1972, *32,* 1029–1033.

Bouchard, T. J., Jr., & McGee, M. G. Sex differences in human spatial ability: Not an X-linked recessive gene effect. *Social Biology,* 1977, *24,* 332–335.

Brinkmann, E. H. Programmed instruction as a technique for improving spatial visualization. *Journal of Applied Psychology,* 1966, *50,* 179–184.

Broadbent, D. E. The role of auditory localization in attention and memory span. *Journal of Experimental Psychology,* 1954, *47,* 191–196.

Broverman, D. M., & Klaiber, E. L. Negative relationships between abilities. *Psychometrika,* 1969, *34,* 5–20.

Broverman, D. M., Klaiber, E. L., Kobayashi, Y., & Vogel, W. Roles of activation and inhibition in sex differences in cognitive abilities. *Psychological Review,* 1968, *75,* 23–50.

Brown, F. R. *The effect of an experimental course in geometry on ability to visualize in three dimensions.* Unpublished doctoral dissertation, Columbia University, 1954.

Brown, W., & Stephenson, W. A test of the theory of two factors. *British Journal of Psychology,* 1933, *23,* 352–370.

Bryden, M. P. Left–right differences in tachistoscopic recognition: Directional scanning or cerebral dominance? *Perceptual and Motor Skills,* 1966, *23,* 1127–1134.

Buffery, A. W. H. Evidence for the asymmetrical lateralization of cerebral function. *Bulletin of the British Psychological Society,* 1968, *21,* 29.

Buffery, A. W. H. Sex differences in the development of hand preference, cerebral dominance for speech and cognitive skill. *Bulletin of the British Psychological Society,* 1970, *23,* 233.

Buffery, A. W. H. An automated technique for the study of the development of cerebral mechanisms subserving linguistic skill. *Proceedings of the Royal Society of Medicine,* 1971, *64,* 919–922. (a)

Buffery, A. W. H. Sex differences in the development of cognitive skills. *Bulletin of the British Psychological Society,* 1971, *24,* 242–243. (b)

Buffery, A. W. H. Sex differences in the development of hemispheric asymmetry of function in the human brain. *Brain Research,* 1971, *31,* 364–365. (c)

Buffery, A. W. H., & Gray, J. H. Sex differences in the development of spatial and linguistic skills. In C. Ounstead & D. C. Taylor (Eds.), *Gender differences: Their ontogeny and significance.* London: Churchill, 1972.

Churchill, B. D., Curtis, J. M., Coombs, C. H., & Hassell, T. W. Effect of engineer school training on the Surface Development Test. *Educa-*

tional Psychological Measurement, 1942, *2*, 279–280.

Clarke, G. *The range and nature of factors in perceptual tests*. Unpublished doctoral dissertation, University of London, 1936.

Cooper, L. A., & Shepard, R. W. Mental transformations in the identification of left and right hands. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, *104*, 48–56.

Corah, N. L. Differentiation of children and their parents. *Journal of Personality*, 1965, *33*, 300–308.

Corkin, S. Tactually-guided maze learning in man: Effects of unilateral cortical excisions and bilateral hippocampal lesions. *Neuropsychologia*, 1965, *3*, 339–351.

Costello, C. G. The control of visual imagery in mental disorders. *Journal of Mental Science*, 1957, *103*, 840–849.

Cowley, J. J., & Griesel, R. D. The development of second-generation low protein rats. *Journal of Genetic Psychology*, 1963, *103*, 233–242.

Cox, J. W. *Mechanical aptitude*. London: Methuen, 1928.

Curry, F. K. W. A comparison of left-handed and right-handed subjects on verbal and non-verbal dichotic listening tasks. *Cortex*, 1967, *3*, 343–352.

Dawson, J. L. M. Cultural and physiological influences upon spatial-perceptual process in West Africa: Part I. *International Journal of Psychology*, 1967, *2*, 115–128. (a)

Dawson, J. L. M. Cultural and physiological influences upon spatial-perceptual processes in West Africa: Part II. *International Journal of Psychology*, 1967, *2*, 171–185. (b)

Dawson, J. L. M. Effects of sex hormones on cognitive style in rats and men. *Behavior Genetics*, 1972, *2*, 21–42.

DeFries, J. C., et al. Near identity of cognitive structure in two ethnic groups. *Science*, 1974, *183*, 338–339.

DeFries, J. C., et al. Parent–offspring resemblance for specific cognitive abilities in two ethnic groups. *Nature*, 1976, *261*, 131–133.

Dempster, J. J. B. Symposium on the selection of pupils for different types of secondary schools: V. The selector's point of view. *British Journal of Educational Psychology*, 1948, *18*, 121–133.

Drew, L. J. An investigation into the measurement of technical ability. *Occupational Psychology*, 1944, *21*, 34–48.

Ehrlichman, H. I. Hemispheric functioning and individual differences in cognitive abilities (Doctoral dissertation, New School for Social Research, 1971). *Dissertation Abstracts International*, 1972, *33*, 2319B. (University Microfilms No. 72-27,869)

Eisenberg, T. A., & McGinty, R. L. On spatial visualization in college students. *Journal of Psychology*, 1977, *95*, 99–104.

Emmett, W. G. Evidence of a space factor at 11 plus and earlier. *British Journal of Psychology*, 1949, *2*, 3–16.

Estes, S. G. A study of five tests of spatial ability. *Journal of Psychology*, 1942, *13*, 265–271.

Fagan-Dubin, L. Lateral dominance and development of cerebral specialization. *Cortex*, 1974, *10*, 69–74.

Faubian, R. W., Cleveland, E. A., & Hassell, T. W. The influence of training on mechanical aptitude test scores. *Educational Psychological Measurement*, 1942, *2*, 91–94.

Flaugher, F. L., & Rock, D. A. Patterns of ability factors among four ethnic groups. *Proceedings of the 80th Annual Convention of the American Psychological Association*, 1972, *7*, 27–28. (Summary)

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs* (No. 5). Chicago: University of Chicago Press, 1951.

Fruchter, B. Measurement of spatial abilities: History and background. *Educational and Psychological Measurement*, 1954, *14*, 387–395.

Galton, F. Statistics of mental imagery. *Mind*, 1880, *5*, 300–318.

Galton, F. *Inquiries into human faculty and its development*. London: Macmillan, 1883.

Garai, J. E., & Scheinfeld, A. Sex differences in mental and behavioral traits. *Genetic Psychology Monographs*, 1968, *77*, 169–299.

Gardner, R. W., Jackson, D. N., & Messick, S. J. Personality organization in cognitive controls and intellectual abilities. *Psychological Issues*, 1960, *2*(Whole No. 8).

Garron, D. C. Sex-linked recessive inheritance of spatial and numerical abilities and Turner's syndrome. *Psychological Review*, 1970, *77*, 147–152.

Garron, D. C. Intelligence among persons with Turner's syndrome. *Behavior Genetics*, 1977, *7*, 105–127.

Geiringer, E. R., & Hyde, J. J. Sex differences on Piaget's water level task: Spatial incognito. *Perceptual and Motor Skills*, 1976, *42*, 1323–1328.

Geschwind, N. The anatomical basis of hemispheric differentiation. In S. J. Dimond & J. G. Beaumont (Eds.), *Hemisphere function in the human brain*. New York: Wiley, 1974.

Ghiselli, E. E. *The validity of occupational aptitude tests*. New York: Wiley, 1966.

Ghiselli, E. E. The validity of aptitude tests in personnel selection. *Personnel Psychology*, 1973, *26*, 461–477.

Gibson, E. J. Improvement in perceptual judgments as a function of controlled practice and training. *Psychological Bulletin*, 1953, *50*, 401–431.

Goldberg, J., & Meredith, W. A longitudinal study of spatial ability. *Behavior Genetics*, 1975, *5*, 127–135.

Goldstein, A. G., & Chance, J. E. Effects of practice on sex-related differences in performance on embedded figures. *Psychonomic Science*, 1965, *3*, 361–362.

Goodenough, D. R., & Eagle, C. T. A modification of the Embedded-Figures Test use with young children. *Journal of Genetic Psychology*, 1963,

103, 67–74.

Goodenough, D. R., et al. A study of X chromosome linkage with field dependence and spatial visualization. *Behavior Genetics*, 1977, *7*, 373–387.

Goodglass, H., & Quadfasel, F. A. Language laterality in left-handed aphasics. *Brain*, 1954, *77*, 521–548.

Gordon, R. An investigation into some of the factors that favor the formation of stereotyped images. *British Journal of Psychology*, 1949, *39*, 156–167.

Gough, H., & Olton, R. M. Field independence as related to nonverbal measures of perceptual performance and cognitive ability. *Journal of Consulting and Clinical Psychology*, 1972, *38*, 338–342.

Graves, M. F., & Koziol, S. Noun plural development in primary grade children. *Child Development*, 1971, *42*, 1165–1173.

Gross, F. The role of set in perception of the upright. *Journal of Personality*, 1959, *27*, 95–103.

Guilford, J. P., Fruchter, B., & Zimmerman, W. S. Factor analysis of the Army Air Forces, Sheppard Field Battery of Experimental Aptitude Tests. *Psychometrika*, 1952, *16*, 45–68.

Guilford, J. P., & Lacey, J. I. *Printed Classification Tests, A.A.F.* (Aviation Psychological Progress Research Rep. No. 5). Washington, D.C.: U.S. Government Printing Office, 1947.

Guilford, J. P., & Zimmerman, W. S. Some A.A.F. findings concerning aptitude factors. *Occupations*, 1947, *26*, 154–159.

Guilford, J. P., & Zimmerman, W. S. *Guilford-Zimmerman Aptitude Survey*. Orange, Calif.: Sheridan Psychological Services, 1953.

Guttman, R. Genetic analysis of analytical spatial ability: Raven's progressive matrices. *Behavior Genetics*, 1974, *4*, 273–284.

Hakstian, A. R., & Cattell, R. B. The checking of primary ability structure on a broader basis of performances. *British Journal of Educational Psychology*, 1974, *44*, 140–154.

Hamley, H. R. *The testing of intelligence*. London: Evans, 1935.

Hardyck, C., & Petrinovich, L. F. Left-handedness. *Psychological Bulletin*, 1977, *84*, 385–404.

Harris, L. J. Sex differences in spatial ability: Possible environmental, genetic, and neurological factors. In M. Kinsbourne (Ed.), *Asymmetrical function of the brain*. New York: Cambridge University Press, 1978.

Harshman, R. A., & Remington, R. Sex language, and the brain: Part I. A review of the literature on adult sex differences in lateralization. *UCLA Working Papers in Phonetics*, 1976, *31*, 86–103.

Hartlage, L. C. Sex-linked inheritance of spatial ability. *Perceptual and Motor Skills*, 1970, *31*, 610.

Hermelin, B., & O'Connor, N. Functional asymmetry in the reading of Braille. *Neuropsychologia*, 1971, *9*, 431–435. (a)

Hermelin, B., & O'Connor, N. Right and left handed reading of Braille. *Nature*, 1971, *231*, 470. (b)

Heston, J. C. A factor analysis of some clinical performance tests. *Journal of Applied Psychology*, 1943, *27*, 135–139.

Hills, J. R. Factor analyzed abilities and success in college mathematics. *Educational Psychological Measurement*, 1957, *17*, 615–622.

Hogben, L. Filial and fraternal correlations in sex-linked inheritance. *Proceedings of the Royal Society of Edinburgh*, 1932, *52*, 331–336.

Holliday, F. An investigation into selection of apprentices for the engineering industry. *Occupational Psychology*, 1940, *14*, 69–81.

Holliday, F. The relations between psychological test scores and subsequent proficiency of apprentices in the engineering industry. *Occupational Psychology*, 1943, *17*, 168–185.

Holzinger, K. J., & Swineford, F. The relation of two bi-factors to achievement in geometry and other subjects. *Journal of Educational Psychology*, 1946, *37*, 257–265.

Hubbert, H. B. The effect of age on habit formation in the albino rat. *Behavior Monographs*, 1915, *2*, 1–55.

Humphreys, L. G., & Taber, T. Ability factors as a function of advantaged and disadvantaged groups. *Journal of Educational Measurement*, 1973, *10*, 107–115.

Hunt, D., & Randhawa, B. S. Relationship between and among cognitive variables and achievement in computational science. *Educational and Psychological Measurement*, 1973, *33*, 921–928.

Hunter, R. S. Aptitude tests for the machine shop. *Industrial Arts and Vocational Education*, 1945, *34*, 58–64.

Hyde, J. S., Geiringer, E. R., & Yen, W. On the empirical relation between spatial ability and sex differences in other aspects of cognitive performance. *Multivariate Behavioral Research*, 1975, *10*, 289–301.

James, W. E., Mefferd, R. B., & Wieland, B. Repetitive psychometric measures: Handedness and performance. *Perceptual and Motor Skills*, 1967, *25*, 209–212.

Jensen, A. R. A theoretical note on sex linkage and race differences in spatial visualization ability. *Behavior Genetics*, 1975, *5*, 151–164.

Johnson, R. G. Assortative marriage for specific cognitive abilities in two ethnic groups. *Human Biology*, 1976, *48*, 343–352.

Karlins, M., Schuerkoff, C., & Kaplan, M. Some factors related to architectural creativity in graduating architecture students. *Journal of General Psychology*, 1969, *81*, 203–215.

Kato, N. A fundamental study of rod and frame test. *Japanese Psychological Research*, 1965, *7*, 61–68.

Kelley, T. L. *Crossroads in the mind*. Stanford, Calif.: Stanford University Press, 1928.

Keogh, B. K., & Ryan, S. R. Use of three measures and field organization with young children. *Perceptual and Motor Skills*, 1971, *33*, 466.

Kershner, J. R., & King, A. J. Laterality of cognitive functions in achieving hemiplegic children.

Perceptual and Motor Skills, 1974, *39,* 1283–1289.

Kimura, D. Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology,* 1961, *15,* 166–171.

Kimura, D. Speech lateralization in young children as determined by an auditory test. *Journal of Comparative and Physiological Psychology,* 1963, *56,* 899–902.

Kimura, D. Left-right differences in the perception of melodies. *Quarterly Journal of Experimental Psychology,* 1964, *16,* 355–358.

Kimura, D. Dual functional asymmetry of the brain in visual perception. *Neuropsychologia,* 1966, *4,* 275–285.

Kimura, D. Functional asymmetry of the brain in dichotic listening. *Cortex,* 1967, *3,* 163–178.

Kimura, D. Spatial localization in left and right visual fields. *Canadian Journal of Psychology,* 1969, *23,* 455–458.

Kimura, D. The asymmetry of the human brain. *Scientific American,* 1973, *228*(3), 70–78.

Klaiber, E. L., Broverman, D. M., & Kobayashi, Y. The automatization cognitive style, androgens, and monoamine oxidase (MAO). *Psycholopharmacologia,* 1967, *11,* 320–336.

Knight, W. H. *A study of the Middlesex Junior Technical School entrance examination.* Unpublished master's thesis, University of London, 1949.

Knox, C., & Kimura, D. Cerebral processing of non-verbal sounds in boys and girls. *Neuropsychologia,* 1970, *8,* 227–238.

Kohs, S. C. *Intelligence measurement.* New York: Macmillan, 1923.

Koussy, A. A. H. The visual perception of space. *British Journal of Psychology Monograph Supplement,* 1935, No. 20.

Kutas, J., McCarthy, G., & Donchin, E. Differences between sinistrals' and dextrals' ability to infer a whole from its parts: A failure to replicate. *Neuropsychologia,* 1975, *13,* 455–464.

Langhorne, M. C. The effects of maze rotation on learning. *Journal of General Psychology,* 1948, *38,* 191–205.

Lansdell, H. The effect of neurosurgery on a test of proverbs. *American Psychologist,* 1961, *16,* 448.

Lansdell, H. A sex difference in effect of temporal lobe neurosurgery on design preference. *Nature,* 1962, *194,* 852–854.

Lansdell, H. Effect of extent of temporal lobe ablations on two lateralized deficits. *Physiology and Behavior,* 1968, *3,* 271–273. (a)

Lansdell, H. The use of factor scores from the Wechsler-Bellevue Scale of Intelligence in assessing patients with temporal lobe removals. *Cortex,* 1968, *4,* 257–268. (b)

Lansdell, H. Effect of neurosurgery on the ability to identify popular word associations. *Journal of Abnormal Psychology,* 1973, *81,* 255–258.

Lansdell, H., & Davie, J. C. Massa intermedia: Possible relation to intelligence. *Neuropsychologia,* 1972, *10,* 207–210.

Leff, G. *Technical aptitude in boys and girls aged 12 years with special reference to Alexander's Performance Scale.* Unpublished master's thesis, University of London, 1949.

Levy, J. Possible basis for the evolution of lateral specialization of the human brain. *Nature,* 1969, *224,* 614–615.

Levy, J. Cerebral lateralization and spatial ability. *Behavior Genetics,* 1976, *6,* 171–188. (a)

Levy, J. A review of evidence for a genetic component in the determination of handedness. *Behavior Genetics,* 1976, *6,* 429–453. (b)

Likert, R., & Quasha, W. H. *Manual for the Revised Minnesota Paper Form Board Test.* New York: The Psychological Corporation, 1970.

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. *Race differences in intelligence.* San Francisco: Freeman, 1975.

Loehlin, J. C., Sharan, S., & Jacoby, R. In pursuit of the "spatial gene": A family study. *Behavior Genetics,* 1978, *8,* 27–41.

Maccoby, E. E. Sex differences in intellectual functioning. In E. E. Maccoby (Ed.), *The development of sex differences.* Stanford, Calif.: Stanford University Press, 1966.

Maccoby, E., & Jacklin, C. N. *The psychology of sex differences.* Stanford, Calif.: Stanford University Press, 1974.

Marjoribanks, K. Environment, social class, and mental abilities. *Journal of Educational Psychology,* 1972, *63,* 103–109.

Marks, D. F. Individual differences in the vividness of visual imagery and their effect on function. In P. W. Sheehan (Ed.), *The function and nature of imagery.* New York: Academic Press, 1972.

Martin, G. C. Test batteries for auto mechanics and apparel design. *Journal of Applied Psychology,* 1951, *35,* 20–22.

Mather, K., & Jinks, J. L. Correlation between relatives arising from sex-linked genes. *Nature,* 1963, *198,* 314–315.

Mattson, M. L. The relation between the habit to be acquired and the form of the learning curve in young children. *Genetic Psychology Monographs,* 1933, *13,* 299–398.

McFarlane, M. A study of practical ability. *British Journal of Psychology Monograph Supplement,* 1925, *8*(1, Whole No. 8).

McFie, J., Piercy, M. F., & Zangwill, O. L. Visual-spatial agnosia associated with lesions of the right cerebral hemisphere. *Brain,* 1950, *83,* 243–260.

McGee, M. G. Laterality, hand preference, and human spatial ability. *Perceptual and Motor Skills,* 1976, *42,* 781–782.

McGee, M. G. A family study of human spatial abilities (Doctoral dissertation, University of Minnesota, Minneapolis, 1976). *Dissertation Abstracts International,* 1977, *37,* 6396. (University Microfilms No. 77-12,836)

McGee, M. G. Effects of training and practice on sex differences in Mental Rotation Test scores. *Journal of Psychology,* 1978, *100,* 87–90. (a)

McGee, M. G. Handedness and mental rotation. *Perceptual and Motor Skills*, 1978, *47*, 641–642. (b)

McGee, M. G. Intrafamilial correlations and heritability estimates for spatial ability in a Minnesota sample. *Behavior Genetics*, 1978, *8*, 77–80. (c)

McGinnis, E. The acquisition and interference of motor habits in young children. *Genetic Psychology Monographs*, 1929, *6*, 207–311.

McGlone, J., & Davidson, W. The relation between cerebral speech laterality and spatial ability with special reference to sex and hand preference. *Neuropsychologia*, 1973, *11*, 105–113.

McGlone, J., & Kertesz, A. Sex differences in cerebral processing of visuo-spatial tasks. *Cortex*, 1973, *9*, 313–320.

McKusick, V. A. *On the X chromosome of man*. Washington, D.C.: American Institute of Biological Sciences, 1964.

McNemar, Q. *The revision of the Stanford-Binet Scale: An analysis of the standardization data*. Boston: Houghton-Mifflin, 1942.

McNemar, Q., & Stone, C. P. The sex difference in rats on three learning tasks. *Journal of Comparative Psychology*, 1932, *14*, 171–180.

Mendicino, L. Mechanical reasoning and space perception: Native capacity or experience. *Personnel and Guidance Journal*, 1958, *36*, 335–338.

Metzler, J., & Shepard, R. N. Rotation of tri-dimensional objects. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium*. New York: Wiley, 1974.

Michael, W. B. Factor analysis of tests and criteria: A comparative study of two AAF pilot populations. *Psychological Monographs*, 1949, *63* (1, Whole No. 298).

Michael, W. B., Guilford, J. P., Fruchter, B., & Zimmerman, W. S. The description of spatial-visualization abilities. *Education and Psychological Measurement*, 1957, *17*, 185–199.

Miller, E. Handedness and the patterns of human ability. *British Journal of Psychology*, 1971, *62*, 111–112.

Milner, B. Laterality effects in audition. In V. B. Mountcastle (Ed.), *Interhemispheric relations and cerebral dominance*. Baltimore, Md.: Johns Hopkins University Press, 1962.

Milner, B. Visually-guided maze learning in man: Effects of bilateral hippocampal, bilateral frontal, and unilateral cerebral lesions. *Neuropsychologia*, 1965, *3*, 317–338.

Milner, B., Taylor, L. B., & Sperry, R. W. Lateralized suppression of dichotically-presented digits after commissural section in man. *Science*, 1968, *161*, 184–185.

Money, J. Cytogenetic and psychosexual incongruities with a note on space-form blindness. *American Journal of Psychiatry*, 1963, *119*, 820–827.

Money, J., Alexander, D., & Walker, H. T., Jr. *A standardized test of direction sense*. Baltimore, Md.: Johns Hopkins University Press, 1965.

Money, J., & Ehrhardt, A. A. *Man and women: Boy and girl*. Baltimore, Md.: Johns Hopkins University Press, 1972.

Money, J., & Granoff, D. IQ and the somatic stigmata of Turner's syndrome. *American Journal of Mental Deficiency*, 1965, *70*, 69–77.

Morf, M. E., Kavanaugh, R. D., & McConville, M. Intratest and sex differences on a portable rod-and-frame test. *Perceptual and Motor Skills*, 1971, *32*, 727–733.

Morris, C. M. A critical analysis of certain performance tests. *Journal of Genetic Psychology*, 1939, *54*, 85–105.

Murphy, L. W. The relation between mechanical ability tests and verbal and non-verbal intelligence tests. *Journal of Psychology*, 1936, *2*, 353–366.

Nash, S. C. *Conceptions and concomitants of sex-role stereotyping*. Unpublished doctoral dissertation, Columbia University, 1973.

Nebes, R. Handedness and the perception of part–whole relationships. *Cortex*, 1971, *7*, 350–356.

Nebes, R. D. Hemispheric specialization in commissurotomized man. *Psychological Bulletin*, 1974, *81*, 1–14.

Nebes, R. D., & Briggs, G. C. Handedness and the retention of visual material. *Cortex*, 1974, *10*, 209–214.

Neimark, E. D., & Santa, J. L. Thinking and concept attainment. *Annual Review of Psychology*, 1975, *26*, 173–205.

Newcombe, F., & Ratcliff, G. Handedness, speech lateralization and ability. *Neuropsychologia*, 1973, *11*, 399–407.

Nichols, P. L. The effects of heredity and environment on intelligence test performance in 4 and 7 year white and negro sibling pairs (Doctoral dissertation, University of Minnesota, Minneapolis, 1970). *Dissertation Abstracts International*, 1971, *32*, 101B. (University Microfilms No. 71-18, 874)

Nunnally, J. C. *Psychometric theory* (2nd ed.). New York: McGraw-Hill, 1978.

O'Connor, J. *Structural visualization*. Boston: Human Engineering Laboratory, 1943.

Okonji, M. O. The differential effects of rural and urban upbringing on the development of cognitive styles. *International Journal of Psychology*, 1969, *4*, 293–305.

Osborne, R. T., & Gregor, A. J. Racial differences in heritability estimates for tests of spatial ability. *Perceptual and Motor Skills*, 1968, *27*, 735–739.

Oxlade, M. N. Further experience with selection tests for power sewing machine operators. *Bulletin of Industrial Psychology and Personnel Practice*, 1951, *8*, 1.

Park, J., et al. Parent–offspring resemblance for specific cognitive abilities in Korea. *Behavior Genetics*, 1978, *8*, 43–52.

Paterson, D. G., Elliott, R. M., Anderson, L. D., Toops, H. A., & Heidbreder, E. *Minnesota mechanical ability tests*. Minneapolis: University of Minnesota Press, 1930.

Petersen, A. C. Physical androgyny and cognitive functioning in adolescence. *Developmental Psychology*, 1976, *12*, 524–533.

Piaget, J., & Inhelder, B. *The child's concept of space*. New York: Humanities Press, 1956.

Pizzamiglio, L., & Cecchini, M. Development of the hemispheric dominance in children from 5 to 10 years of age and their relations with development of cognitive processes. *Brain Research*, 1971, *31*, 363.

Podell, J. E., & Phillips, L. A developmental analysis of cognition as observed in dimensions of Rorschach and objective test performance. *Journal of Personality*, 1959, *27*, 439–463.

Porteus, S. D. The measurement of intelligence: 653 children examined by the Binet and Porteus tests. *Journal of Educational Psychology*, 1918, *9*, 13–31.

Porteus, S. D. *Porteus maze tests: Fifty years' application*. Palo Alto, Calif.: Pacific Books, 1965.

Price, E. J. J. The nature of the practical factor (F). *British Journal of Psychology*, 1940, *30*, 341–351.

Pylyshyn, Z. W. What the mind's eye tells the brain. A critique of mental imagery. *Psychological Bulletin*, 1973, *80*, 1–24.

Ranucci, E. R. *The effect of the study of solid geometry on certain aspects of space percepion abilities*. Unpublished doctoral dissertation, Columbia University, 1952.

Richardson, A. *Mental imagery*. London: Routledge & Kegan Paul, 1969.

Richardson, A. Voluntary control of the memory image. In P. W. Sheehan (Ed.), *The function and nature of imagery*. New York: Academic Press, 1972.

Riddle, M., & Roberts, A. H. Delinquency, delay of gratification, recidivism, and the Porteus maze tests. *Psychological Bulletin*, 1977, *84*, 417–425.

Roff, M. A. A factorial study of tests in the perceptual area. *Psychometric Monographs* (No. 2). Chicago: University of Chicago Press, 1952.

Rudel, R. G., Denckla, M. B., & Spalten, E. The functional asymmetry of Braille letter learning in normal sighted children. *Neurology*, 1973, *24*, 733–738.

Saad, L. G., & Storer, W. O. *Understanding in mathematics*. Edinburgh, Scotland: Oliver & Boyd, 1960.

Saarni, C. I. Piagetian operations and field independence as factors in children's problem-solving performance. *Child Development*, 1973, *44*, 338–345.

Sadownikova-Koltzova, M. P. Genetic analysis of temperment of rats. *Journal of Experiment Zoology*, 1926, *45*, 301–318.

Salkind, N. J. A cross-dimensional study of spatial visualization in young children. *Journal of Genetic Psychology*, 1976, *129*, 339–340.

Santos, J. F., & Murphy, G. An odyssey in perceptual learning. *Bulletin of the Menninger Clinic*, 1960, *24*, 6–17.

Scarr-Salapatek, S. Race, social class, and IQ. *Science*, 1971, *174*, 1285–1295.

Schwartz, D. W., & Karp, S. A. Field dependence in a geriatric population. *Perceptual and Motor Skills*, 1967, *24*, 495–504.

Searleman, A. A review of right hemisphere linguistic capabilities. *Psychological Bulletin*, 1977, *84*, 503–528.

Serra, A., Pizzamiglio, L., Boari, A., & Spera, S. A comparative study of cognitive traits in human sex chromosome aneuploids and sterile and fertile euploids. *Behavior Genetics*, 1978, *8*, 143–154.

Shankweiler, D. Effects of temporal-lobe damage on perception of dichotically presented melodies. *Journal of Comparative and Physiological Psychology*, 1966, *62*, 115–119.

Sheehan, P. W. Reliability of a short test of imagery. *Perceptual and Motor Skills*, 1967, *25*, 744. (a)

Sheehan, P. W. A shortened form of Bett's questionnaire upon mental imagery. *Journal of Clinical Psychology*, 1967, *23*, 386–389. (b)

Shepard, R. N., & Chipman, S. Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1970, *1*, 1–17.

Shepard, R. N., & Feng, C. A chronometric study of mental paper folding. *Cognitive Psychology*, 1972, *3*, 228–243.

Shepard, R. N., & Metzler, J. Mental rotation of three-dimensional objects. *Science*, 1971, *171*, 701–703.

Sherman, J. A. Problem of sex differences in space perception and aspects of intellectual functioning. *Psychological Review*, 1967, *74*, 290–299.

Sherman, J. A. *On the psychology of women: A survey of empirical studies*. Springfield, Ill.: Charles C Thomas, 1971.

Sherman, J. A. Cognitive performance as a function of sex and handedness: An evaluation of the Levy hypothesis. *Psychology of Women Quarterly*, in press.

Shuttleworth, C. W. Tests of technical aptitude. *Occupational Psychology*, 1942, *16*, 175–182.

Silbert, A., Wolff, P. H., & Lilienthal, J. Spatial and temporal processing in patients with Turner's syndrome. *Behavior Genetics*, 1977, *7*, 11–21.

Silverman, A., Adevai, G., & McGough, W. Some relationships between handedness and perception. *Journal of Psychosomatic Research*, 1966, *10*, 151–158.

Silverman, J., Bucksbaum, M., Stierlin, H. Sex differences in perceptual differentiation and stimulus intensity control. *Journal of Personality and Social Psychology*, 1973, *25*, 309–318.

Slater, P. Some group tests of spatial judgment or practical ability. *Occupational Psychology*, 1940, *14*, 40–55.

Slater, P. Tests for selecting secondary and technical school children. *Occupational Psychology*, 1941, *15*, 10.

Smith, I. M. The form-perception factor. A study of a variety of form-relations tests of the existence of a factor of form perception and its relation to success in school subjects (Bachelor of

education thesis, University of Glasgow, 1937). *British Journal of Educational Psychology*, 1938, *3*, 1. (Summary)

Smith, I. M. Measuring spatial abilities in school pupils. *Occupational Psychology*, 1948, *22*, 150–159.

Smith, I. M. *Spatial ability: Its educational and social significance*. London: University of London, 1964.

Sparks, R., & Geschwind, N. Dichotic listening in man after section of neocortical commissures. *Cortex*, 1968, *4*, 3–16.

Sperry, R. W. Hemisphere deconnection and unity in conscious awareness. *American Psychologist*, 1968, *23*, 723–733.

Sperry, R. W. Lateral specialization of cerebral function in surgically separated hemispheres. In F. J. McGuigan & R. A. Schoonover (Eds.), *The psychophysiology of thinking*. New York: Academic Press, 1973.

Sperry, R. W., Gazzaniga, M. S., & Bogen, J. E. Interhemispheric relationships: The neocortical commissures, syndromes of hemisphere disconnection. In P. J. Vinken & G. W. Bruyn (Eds.), *Handbook of clinical neurology* (Vol. 4). New York: North-Holland, 1969.

Spreen, O., Benton, A. L., & Fincham, R. W. Auditory agnosia without aphasia. *Archives of Neurology*, 1965, *13*, 84–92.

Stafford, R. E. Sex differences in spatial visualization as evidence of sex-linked inheritance. *Perceptual and Motor Skills*, 1961, *13*, 428.

Stenquist, J. L. *Mechanical aptitude tests*. New York: World Book, 1922.

Stuart-Mason, A. *Health and hormones*. Harmondsworth, England: Penguin Books, 1963.

Studdert-Kennedy, M., & Shankweiler, D. Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America*, 1970, *48*, 579–594.

Swineford, F. A study in factor analysis. The nature of the general, verbal, and spatial bi-factors. *Supplementary Educational Monograph* (No. 67). Chicago: University of Chicago Press, 1948.

Tanner, J. M. Growth and endocrinology of the adolescent. In L. I. Gardner (Ed.), *Endocrine and genetic diseases of childhood*. Philadelphia, Pa.: Saunders, 1969.

Teegarden, L. Manipulative performance of young adult applicants at a public employment office: Part II. *Journal of Applied Psychology*, 1942, *26*, 754–769.

Thomas, H., Jamison, W., & Hummel, D. D. Observation is insufficient for discovering that the surface of still water is invariantly horizontal. *Science*, 1973, *181*, 173–174.

Thurstone, L. L. *Primary mental abilities*. Chicago: University of Chicago Press, 1938.

Thurstone, L. L. *A factorial study of perception*. Chicago: University of Chicago Press, 1944.

Thurstone, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.

Thurstone, L. L., & Thurstone, T. G. *The Primary Mental Abilities Tests*. Chicago: Science Research Associates, 1941.

Thurstone, L. L., & Thurstone, T. G. *Primary mental abilities technical report*. Chicago: Science Research Associates, 1965.

Tomlin, M. I., & Stone, C. P. Sex differences in learning abilities of albino rats. *Journal of Comparative Psychology*, 1933, *16*, 207–229.

Tryon, R. C. Studies in individual differences in maze ability: I. The determination of individual differences by age, weight, sex, and pigmentation. *Journal of Comparative Psychology*, 1931, *12*, 1–22.

Tuddenham, R. D. A Piagetian test of cognitive development. In W. B. Dockrell (Ed.), *On intelligence: The Toronto symposium on intelligence*. London: Methuen, 1970.

Tyler, L. *The psychology of human differences* (3rd ed.). New York: Appleton-Century-Crofts, 1965.

U.S. Employment Service. *Estimates of worker trait requirements for 4,000 jobs*. Washington, D.C.: U.S. Government Printing Office, 1957.

Vandenberg, S. G. Twin data in support of the Lyon hypothesis. *Nature*, 1962, *194*, 505–506.

Vandenberg, S. G. Hereditary factors in psychological variables in man with special emphasis on cognition. In J. N. Spuhler (Ed.), *Genetic diversity and behavior*. Chicago: Aldine, 1967.

Vandenberg, S. G. The nature and nurture of intelligence. In D. C. Glass (Ed.), *Genetics*. New York: Rockefeller, 1968.

Vandenberg, S. G. A twin study of spatial ability. *Multivariate Behavioral Research*, 1969, *4*, 273–294.

Vandenberg, S. G. The genetics of intelligence. In L. C. Deighton (Ed.), *Encyclopedia of education*. New York: Macmillan, 1971.

Vandenberg, S. G. Sources of variance in performance on spatial tests. In J. Eloit & N. J. Salkind (Eds.), *Children's spatial development*. Springfield, Ill.: Charles C Thomas, 1975.

Vandenberg, S. G., & Kuse, A. R. Mental Rotations: A group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 1978, *47*, 599–604.

Vandenberg, S. G., Stafford, R. E., & Brown, A. M. The Louisville twin study. In S. G. Vandenberg (Ed.), *Progress in human behavior genetics*. Baltimore, Md.: Johns Hopkins University Press, 1968.

Van Voorhis, W. R. *The improvement of space perception ability by training*. Unpublished doctoral dissertation, Pennsylvania State University, 1941.

Vaught, G. M. The relationship of role identification and ego strength on sex differences in the rod-and-frame test. *Journal of Personality*, 1965, *33*, 271–283.

Vignolo, L. A. Auditory agnosia: A review and report of recent evidence. In A. L. Benton (Ed.), *Contribution to clinical neuropsychology*. Chicago: Aldine, 1969.

Waber, D. P. Sex differences in cognition: A function of maturation rate? *Science*, 1976, *192*, 572–574.

Waber, D. P. Sex differences in mental abilities, hemispheric lateralization and rate of physical growth at adolescence. *Developmental Psychology*, 1977, *13*, 29–38.

Wada, J. A. Morphologic asymmetry of human cerebral hemispheres: Temporal and frontal speech zones in 100 adult and 100 infant brains. *Neurology*, 1974, *24*, 349.

Watts, A. F. A group performance test. *Bulletin of the National Foundation for Education Research*, 1953, *2*, 15–21.

Werdelin, I. *Geometrical ability and the space factors in boys and girls*. Lund, Sweden: University of Lund Press, 1961.

Williams, H. S. *Some aspects of the measurement and maturation of mechanical aptitude in boys aged 12 to 14*. Unpublished doctoral dissertation, University of London, 1948.

Williams, T. Family resemblance in abilities: The Wechsler scales. *Behavior Genetics*, 1975, *5*, 405–409.

Witelson, S. F. Sex and the single hemisphere: Specialization of the right hemisphere for spatial processing. *Science*, 1976, *193*, 425–427.

Witelson, S. F., & Pallie, W. Left hemisphere specialization for language in the newborn: Neuroanatomical evidence of asymmetry. *Brain*, 1973, *96*, 641–646.

Witkin, H. A. Individual differences in ease of perception of embedded figures. *Journal of Personality*, 1950, *19*, 1–15.

Witkin, H. A., Dyk, R. B., Faterson, G. E., Goodenough, D. R., & Karp, S. A. *Psychological differentiation*. New York: Wiley, 1962.

Witkin, H. A., Goodenough, D. R., & Karp, S. A. Stability of cognitive style from childhood to young adulthood. *Journal of Personality and Social Psychology*, 1967, *7*, 291–300.

Witkin, H. A., et al. *Personality through perception*. New York: Harper, 1954.

Wittenborn, J. R. Mechanical ability, its nature and measurement: I. An analysis of the variables employed in the preliminary Minnesota experiment. *Educational and Psychological Measurement*, 1945, *5*, 395–409.

Yen, W. M. Sex-linked major-gene influences on selected types of spatial performance. *Behavior Genetics*, 1975, *5*, 281–298.

Zimmerman, W. S. Hypotheses concerning the nature of spatial factors. *Educational and Psychological Measurement*, 1954, *14*, 396–400. (a)

Zimmerman, W. S. The influence of item complexity upon the factor composition of a spatial visualization test. *Educational and Psychological Measurement*, 1954, *14*, 106–119. (b)

Zonderman, A. B., Vandenberg, S. G., Spuhler, K. P., & Fain, P. R. Assortive marriage for cognitive abilities. *Behavior Genetics*, 1977, *7*, 261–271.

Received May 22, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 919–926

Alternatives to Simonton's Analyses of the Interrupted and Multiple-Group Time-Series Designs

James Algina
College of Education
University of Florida

Hariharan Swaminathan
School of Education
University of Massachusetts, Amherst

Statistical procedures for analyzing the interrupted time-series and the multiple-group time-series designs are outlined. The procedures are applicable when several subjects are observed on several pretreatment and posttreatmeant occasions, and the number of subjects is greater than the number of occasions.

Simonton (1977) discussed the use of several linear models for the analysis of data arising in the interrupted time-series design and the multiple-group time-series design. The interrupted time-series design consists of measuring the same subjects on several pretreatment occasions, introducing a treatment, and measuring the subjects on several posttreatment occasions. In the multiple-group time-series design, two groups of subjects are used. In this case, the experimental group is treated in the same manner as the single group in the interrupted time-series design. The control group is measured on the same occasions as the experimental group but does not receive the treatment. Typically, assignment of subjects to the two groups is nonrandom, although the treatment may be assigned randomly to the groups.

The linear models discussed by Simonton (1977) are similar to the univariate polynomial trend analysis model that can be used to estimate regression curves with a quantitative independent variable. However, there is a major difference between the data for which univariate trend analysis is appropriate and the data obtained in the time-series designs. The difference is that to use the univariate trend analysis, different subjects are used at each level of the independent variable. Therefore, it may be reasonable to assume that the residuals from the polynomial trend function are uncorrelated, an assumption that is required for the use of ordinary least squares estimates. The reasonableness of this assumption derives from the fact that each residual characterizes a different subject. Since in the time-series designs the same subjects are measured at each occasion, the residuals in any model for time-series data characterize the same individuals and therefore are likely to be correlated. As a result, ordinary least squares estimates of the parameters may not be efficient and hence should not be used. Simonton recognized this problem and suggested the use of modified generalized least squares estimators. This is a reasonable suggestion.

The analyses discussed by Simonton (1977) consist of estimating the parameters of the linear model and then testing hypotheses about the parameters to examine the hypothesis of a treatment effect. In general, we do not have any objections to the models proposed by Simonton for the interrupted time-series design. Our concern is with the analyses advocated by Simonton, and one purpose of this article is to propose alternative statistical analyses. With regard to the multiple-group time-series design, our position is that the aims of Simonton's analyses can be realized using profile analysis. Therefore, a second purpose of this article is to describe a profile analysis of the data arising from a multiple-group time-series design.

In all of the following, we assume that the number of subjects (N) exceeds the number of occasions (p). Simonton (1977)

Requests for reprints should be sent to James Algina, Foundations of Education, College of Education, University of Florida, Gainesville, Florida 32611.

suggested that N should be larger than p, preferably twice as large. Hence, our criterion for N is not more restrictive than Simonton's.

Critique of Simonton's Analyses

It can be shown that each of Simonton's (1977) models for the interrupted time-series design can be expressed as a special case of

$$\bar{\mathbf{X}} = \mathbf{A}\beta + \mathbf{e} \tag{1}$$

where $\bar{\mathbf{X}}$ is the $(px1)$ sample mean vector, $\mathbf{A}$ is a (pxr) known-design matrix, β is an $(rx1)$ vector of unknown parameters, and $\mathbf{e}$ is a $(px1)$ random vector of residuals. The vector $\bar{\mathbf{X}}$ is calculated from the N realizations, $\mathbf{X}_i$, of the random vector $\mathbf{X}$. The variance–covariance matrix of $\mathbf{X}$ is Σ. If the model in Equation 1 is correct, then the population mean vector $\mu = \mathbf{A}\beta$ and the variance–covariance matrix of $\mathbf{e}$, and hence, $\bar{\mathbf{X}}$, is $\mathbf{V} = \Sigma/N$.

The vector β should not be estimated using the ordinary least squares estimator $\hat{\beta}_1 = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\bar{\mathbf{X}}$, since it may not be efficient if the variance–covariance matrix of $\mathbf{e}$ is not $\sigma^2 I$, (i.e., if the elements of $\mathbf{e}$ are correlated or have unequal variances). An appropriate estimator for β is the modified generalized least squares estimator $\hat{\beta} = (\mathbf{A}'\hat{\mathbf{V}}^{-1}\mathbf{A})^{-1}\mathbf{A}'\hat{\mathbf{V}}^{-1}\bar{\mathbf{X}}$, where $\hat{\mathbf{V}}$ is an estimator of $\mathbf{V}$. The problem in using $\hat{\beta}$ is developing an estimator for $\mathbf{V}$.

As noted above, Simonton (1977) recognized that the elements of $\mathbf{e}$ are likely to be correlated and attempted to use $\hat{\beta}$ by calculating an estimate of $\mathbf{V}$ under the assumption that the elements of $\mathbf{e}$ conform to a first-order stationary autoregressive model. This implies that

$$e_t = \rho e_{t-1} + \delta_t, \quad t = 2, \ldots, p$$

with the mean of $e_t = 0$ and the variance of e_t constant for $t = 1, \ldots, p$. This assumption is unlikely to be correct for two reasons. First, Lord (1963) has shown that the variance of a variable is likely to increase over time. Second, Jöreskog (1970) has pointed out that an autoregressive model is unlikely to fit data perturbed by measurement error.

More important than the likely failure of the data to meet the autoregressive assumption is that it is not necessary to make the assumption to estimate β, provided that $N > p$. Use of $\hat{\beta}$ requires an estimator for $\mathbf{V} = \Sigma/N$. This can be accomplished by using the sample variance-covariance matrix

$$\mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$$

as an estimator of Σ and using $\mathbf{S}/N$ as an estimator of $\mathbf{V}$. This approach requires no assumptions about the stochastic process generating $\mathbf{e}$. The estimator $\hat{\beta}$ then becomes $\hat{\beta} = (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\bar{\mathbf{X}}$.

A second problem with the analyses suggested by Simonton (1977) is that he did not provide a method for evaluating the fit of the model to the data. Therefore, it is possible to estimate the vector β and use this estimate to make an inference about a treatment effect even though the model fits the data badly. A third problem is that Simonton proposed the use of approximate tests of significance to test hypotheses about β. However, exact tests of significance are available when $\mathbf{S}$ is used to estimate Σ, and use of these tests is preferable to use of the approximate tests advocated by Simonton.

With regard to the multiple-group time-series design, our major criticism was discussed previously. The procedures reported by Simonton, and their attendant assumptions, are unnecessary since the aims of his procedures can be accomplished using profile analysis.

Interrupted Time-Series Design

As noted above the models discussed by Simonton (1977) are special cases of the linear model given in Equation 1. The models discussed by Simonton are called changed-level (with permanent, transient, or dampened change level) and changed-slope models. The first purpose of this section is to illustrate how each of these models is expressed in terms of Equation 1. Suppose that the subjects are measured on six occasions, three pretreatment and three posttreatment. The permanent changed-level model is expressed as

$$\begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \\ \bar{X}_5 \\ \bar{X}_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}.$$

The model states that the means for the pretreatment occasions are at the same level. After the treatment, the means move to a new level. The hypothesis that $\beta_1 = 0$ is tested to determine if a treatment effect has occurred. If the hypothesis that $\beta_1 = 0$ cannot be rejected, then the means for the entire time-series design are at the same level and a treatment effect cannot be inferred.

The transient changed-level model can be expressed as

$$\begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \\ \bar{X}_5 \\ \bar{X}_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}.$$

Here the model states that the mean for the occasion immediately following the treatment moves to a level different from the common level of all the other means. For the same reason as with the previous model, the hypothesis that $\beta_1 = 0$ is of interest.

The dampened changed-level model is expressed as

$$\begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \\ \bar{X}_5 \\ \bar{X}_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & .5 \\ 1 & .25 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}.$$

The model states that the means for the pretreatment occasions are all at the same level. After the treatment the mean increases, but over time the means tend to drop back to the initial level. Again, the hypothesis that $\beta_1 = 0$ is tested.

As Simonton (1977) noted, the previous models all assume that the population means would be equal if it were not for the intervening treatment. An alternative assumption is that there is a linear or higher order trend in the population means. Such an assumption can easily be incorporated in any of the changed-level models above. To illustrate, suppose that there is a hypothesized quadratic trend and a dampened level change. The model can be expressed as

$$\begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \\ \bar{X}_5 \\ \bar{X}_6 \end{bmatrix} = \begin{bmatrix} 1 & -5 & 5 & 0 \\ 1 & -3 & -1 & 0 \\ 1 & -1 & -4 & 0 \\ 1 & 1 & -4 & 1 \\ 1 & 3 & -1 & .5 \\ 1 & 5 & 5 & .25 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}.$$

In this model, the parameter of interest is β_3. If the hypothesis that $\beta_3 = 0$ cannot be rejected, then a quadratic curve adequately fits the data, and hence, a treatment effect cannot be inferred. The logic here is that the same trend is evident both in the pretreatment and posttreatment means, so it is difficult to claim that the treatment had an effect. This notion was discussed by Campbell (1963).

The changed-slope model discussed by Simonton (1977) can be expressed as

$$\begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \\ \bar{X}_5 \\ \bar{X}_6 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}.$$

The model states that there is a linear trend in the pretreatment and posttreatment data. With this model there are two null hypotheses to be tested: the hypothesis of a common intercept, H_{01}: $\beta_1 = \beta_3$, and the hypothesis of a common slope, H_{02}: $\beta_2 = \beta_4$. If neither hypothesis is rejected, then the linear trend for both the pretreatment and posttreatment data is the same, and hence, a treatment effect cannot be inferred. If either is rejected, the hypothesis of a treatment effect is supported. With more pretreatment and posttreatment occasions, this model can be generalized to include higher order trends. The analysis of data that used a linear model was discussed by Algina and Swaminathan (1977). The analysis of data that used a curvilinear model was discussed in detail by Swaminathan and Algina (1977). At this point the discussion turns to the statistical procedures for testing the fit of a particular model to the data and for testing hypotheses about the parameters of the model. These procedures are based on an article by Rao (1959) and have been discussed by Swaminathan and Algina.

Let $\epsilon(\ )$ denote the expectation operator. Since $\epsilon(\bar{\mathbf{X}}) = \mu$, it follows that

$$\epsilon(\bar{\mathbf{X}}) = \mu = \mathbf{A}\beta + \epsilon(\mathbf{e}). \tag{2}$$

Let $\mathbf{Q}[(p - r)xp]$ be a nonzero matrix constructed so that $\mathbf{QA} = \mathbf{0}$. The matrix $\mathbf{Q}$ can be chosen as a basis for the matrix $[\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}']$. Premultiplying Equation 2 by $\mathbf{Q}$, we obtain

$$\mathbf{Q}\mu = \mathbf{QA}\beta + \mathbf{Q}\epsilon(\mathbf{e}),$$
$$= \mathbf{Q}\epsilon(\mathbf{e}).$$

As noted above, if the hypothesized model is correct, then $\mu = \mathbf{A}\beta$, and hence $\epsilon(\mathbf{e}) = \mathbf{0}$. Therefore, if the model is correct, $\mathbf{Q}\mu = \mathbf{0}$. On the other hand, if the hypothesized model given by Equation 1 is not correct, then $\mathbf{Q}\mu \neq \mathbf{0}$. Therefore, to test the adequacy of the model, the hypothesis

$$H_0\colon \mathbf{Q}\mu = \mathbf{0} \tag{3}$$

is tested. Failure to reject the hypothesis given by Equation 3 supports the contention that Equation 1 is the correct model. Conversely, rejection of the hypothesis indicates that Equation 1 does not adequately fit the data.

It is well-known that the hypothesis given by Equation 3 may be tested using Hotelling's T^2 statistic,

$$T^2 = N\bar{\mathbf{X}}'\mathbf{Q}'(\mathbf{QSQ}')^{-1}\mathbf{Q}\bar{\mathbf{X}}, \tag{4}$$

where $\bar{\mathbf{X}}$ is the sample mean vector, and $\mathbf{S}$ is the sample variance–covariance matrix. The test statistic may be transformed to the variate

$$F = \frac{N - p + r}{(p - r)(N - 1)} T^2,$$

which is compared to the $(1 - \alpha)$th fractile of the F distributions with $p - r$ and $N - p + r$ degrees of freedom. The matrix $\mathbf{Q}$ is not unique, but Rao (1959) showed that the test statistic is invariant to the choice of $\mathbf{Q}$.

If the test for the fit of the model to the data indicates that the model is inadequate, then a new model must be proposed. If the model adequately fits the data, then the vector β is estimated and the relevant hypothesis is tested. The modified generalized least squares estimate of β is $\hat{\beta} = (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\bar{\mathbf{X}}$.

All of the hypotheses concerning β can be expressed as special cases of the hypothesis $\mathbf{G}\beta = \mathbf{0}$, where $\mathbf{G}$ is a (kxr) matrix of rank $k \leq r$. Rao (1959) has shown that this hypothesis can be tested using the test statistic

$$F = \frac{N(N+r-p-k)\hat{\beta}'\mathbf{G}'[\mathbf{G}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{G}']^{-1}\mathbf{G}\hat{\beta}}{k[(N-1)+T^2]}, \tag{5}$$

where T^2 is the quantity defined in Equation 5. The quantity F follows the F distribution with k and $N - k - p + r$ degrees of freedom. For the changed-level models, $\mathbf{G}$ is a $(1xr)$ vector, say $\mathbf{g}'$, with 0 as the first $r - 1$ elements and 1 as the remaining element. In this case, the test statistic reduces to

$$F = \frac{N(N + r - p - 1)(\mathbf{g}'\hat{\beta})^2}{\mathbf{g}'(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{g}[(N - 1) + T^2]}. \tag{6}$$

The inner product $\mathbf{g}'\hat{\beta}$ simply picks out the estimate of the changed-level parameter from the vector $\hat{\beta}$. The quantity F given by Equation 6 follows the F distribution with 1 and $(N - 1) - (p - r)$ degrees of freedom. If the test statistic exceeds the critical value F, then the hypothesis of a treatment effect is supported.

It should be noted that the test statistic given above has an exact distribution in contrast to the test statistic given by Simonton (1977), which has only an approximate distribution. Hence, with the procedure outlined here, it is possible to control the alpha value precisely.

As noted previously, the changed-slope model discussed by Simonton (1977) is a special case of the class of models discussed by Swaminathan and Algina (1977). Since the statistical details are available in that article, only a verbal description of the analysis will be given. Again the logic of the analysis is based on Campbell's (1963) observation that if a single polynomial adequately fits the entire time-series design, then a treatment effect cannot be inferred. The steps in the Swaminathan-Algina analysis are

1. Test the adequacy of a model which specifies that the pretreatment and posttreatment portions of the series are adequately fit by separate polynomial regression curves of degree $r - 1$. Refer to this model as the complete model.

2. If the complete model is not adequate, test whether an $(r - 1)$th-degree polynomial adequately fits one portion of the time-series

design but not the other. If the answer is affirmative, then different regression curves fit the two portions of the series, and the hypothesis of a treatment effect cannot be rejected.

3. If the answer in Step 2 is negative, raise the degree of the polynomial, and repeat Steps 1 and 2. This cycle is followed until the complete model is accepted or rejected, and the answer in Step 2 is affirmative.

4. In the event the complete model is accepted, estimate the regression coefficients, and test the equality of the coefficients of the pretreatment and posttreatment portions. If the two sets of coefficients are equal, then the same curve fits both portions, and the hypothesis of a treatment effect is not tenable. If the hypothesis of equality is rejected, then the hypothesis of a treatment effect is supported.

Multiple Time-Series Design

The multiple-group time-series design consists of observing an experimental and a control group on several pretreatment occasions, introducing a treatment to the experimental group, and observing the two groups on several posttreatment occasions. In essence, Simonton's (1977) proposed analysis consists of fitting four separate first-degree polynomials: one to the pretreatment means for the experimental group, the second to the posttreatment means for the experimental group, the third to the pretreatment means for the control group, and the fourth to the posttreatment means for the control group. Inferences about treatment effects are based on statistical comparison of the slopes and the intercepts of the various equations. Again, an objection to this kind of analysis is that first-degree polynomials may not adequately fit the four sets of means. However, there is no way to detect such lack of fit from Simonton's analysis. A standard multivariate technique, profile analysis (Morrison, 1967, pp. 141–148), can be used to accomplish the aims of Simonton's analysis of the multiple-group time-series design without making the assumption that polynomials of any degree fit the data.

Simonton (1977) suggests that in an optimal situation subjects would be randomly assigned to experimental and control groups. He recognizes that this may be impossible and suggests that equality of the intercepts and slopes of the pretreatment equations for the experimental and control groups is a necessary condition for the groups to be considered equivalent. If the equality condition fails, he suggests dropping the control group and analyzing the experimental group data as an interrupted time-series experiment. In our view, a more appropriate analysis would begin by testing the hypothesis that there is no Groups × Occasions interaction. If an analysis indicates that there is no interaction, then a treatment effect is not supported. With no interaction, the mean time series for the two groups are parallel. The relationship between the mean time series remains the same over the course of the study, and therefore, it is difficult to argue that a treatment effect has occurred.

Let μ_e and μ_c denote $(px1)$ population mean vectors for the experimental and control groups, respectively. The hypothesis of no Groups × Occasions interaction can be expressed as

$$H_0:\ \mathbf{C}(\mu_e - \mu_c) = \mathbf{0}, \qquad (7)$$

where $\mathbf{0}$ is a $(px1)$ null vector and $\mathbf{C}$ is a $[(p-1)xp]$ matrix,

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}.$$

It is well-known that under the assumption that the experimental and control group observations are each multinormally distributed with equal variance-covariance matrices, the hypothesis given by Equation 7 can be tested using Hotelling's T^2 statistic,

$$T^2 = \frac{N_e N_c}{N_e + N_c}(\bar{\mathbf{X}}_e - \bar{\mathbf{X}}_c)'\mathbf{C}'(\mathbf{CSC}')^{-1} \times \mathbf{C}(\bar{\mathbf{X}}_e - \bar{\mathbf{X}}_c),$$

where N_e and N_c are the numbers of cases in the experimental and control groups, $\bar{\mathbf{X}}_e$ and $\bar{\mathbf{X}}_c$ are $(px1)$ sample mean vectors, and $\mathbf{S}(pxp)$ is the pooled sample variance–covariance matrix. The test statistic can be transformed to the variate

$$F = \frac{N_e + N_c - p}{(N_e + N_c - 2)(p-1)} T^2,$$

which follows the F distribution with $p-1$ and $N_e + N_c - p$ degrees of freedom.

As indicated earlier, if the hypothesis given by Equation 7 is not rejected, the hypothesis of a treatment effect is untenable. If the hypothesis is rejected, no conclusion can be made because an interaction can occur whether or not there is a treatment effect. For example, suppose that an interaction exists for the pretreatment portion of the experimental and control group time series. Then the hypothesis given by Equation 7 should be rejected, but this hardly indicates that there is a treatment effect. Rather, it indicates that because of nonrandom assignment to groups, the mean time series would not have been parallel even if the intervention had not occurred. Now suppose that a Groups × Occasions interaction does *not* exist for the pretreatment means but does exist for the posttreatment means. Here the hypothesis given by Equation 7 should be rejected, and there is evidence supporting the hypothesis of a treatment effect, that is, that although the pretreatment mean time series were parallel, after the treatment the mean time series became nonparallel, which suggests that the treatment had some effect.

The above discussion suggests that if for the entire time series, the hypothesis of no Groups × Occasions interaction is rejected, then the next phase of analysis should determine at which point the deviation or deviations from parallelism occur: on the pretreatment occasions, on the posttreatment occasions, or on both types of occasions. If an interaction occurs only for the posttreatment occasions, then the hypothesis of a treatment effect is supported. If an interaction occurs for the pretreatment occasions, then the control group must be dropped, and the experimental group data should be analyzed as an interrupted time series experiment.

Let μ_{1e}, μ_{1c}, μ_{2e}, and μ_{2c} represent the pretreatment and posttreatment population mean vectors for the experimental and control groups, respectively. Let there be m pretreatment occasions and s posttreatment occasions. The hypothesis of no Groups × Occasions interaction for the pretreatment means may be expressed as

$$H_0\colon\ \mathbf{C}_1(\mu_{1e} - \mu_{1c}) = \mathbf{0}, \qquad (8)$$

where $\mathbf{C}_1$ has the same general form as $\mathbf{C}$ but has dimensions $[(m-1)xm]$. The relevant hypothesis for the posttreatment means may be expressed as

$$H_0\colon\ \mathbf{C}_2(\mu_{2e} - \mu_{2c}) = \mathbf{0}, \qquad (9)$$

where again $\mathbf{C}_2$ has the same general form as $\mathbf{C}$ but has dimensions $[(s-1)xs]$. The test statistics for these two hypotheses are

$$T^2{}_1 = \frac{N_eN_c}{N_e+N_c}(\bar{\mathbf{X}}_{1e} - \bar{\mathbf{X}}_{1c})'\mathbf{C}'_1(\mathbf{C}_1\mathbf{S}_1\mathbf{C}'_1)^{-1} \times \mathbf{C}_1(\bar{\mathbf{X}}_{1e} - \bar{\mathbf{X}}_{1c})$$

and

$$T^2{}_2 = \frac{N_eN_c}{N_e+N_c}(\bar{\mathbf{X}}_{2e} - \bar{\mathbf{X}}_{2c})'\mathbf{C}'_2(\mathbf{C}_2\mathbf{S}_2\mathbf{C}'_2)^{-1} \times \mathbf{C}_2(\bar{\mathbf{X}}_{2e} - \bar{\mathbf{X}}_{2c}),$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are pooled-sample variance–covariance matrices for the pretreatment and posttreatment observations, respectively.

The test statistics can be transformed to the variates

$$F_1 = \frac{N_e + N_c - m}{(N_e + N_c - 2)(m-1)}T^2{}_1,$$

and

$$F_2 = \frac{N_e + N_c - s}{(N_e + N_c - 2)(s-1)}T^2{}_2.$$

The critical value to which F_1 and F_2 are compared is the same as the critical value to which F is compared in testing the hypothesis given by Equation 7. Using this critical value keeps the overall α rate at the nominal level. If the hypothesis given by Equation 8 is not rejected and the hypothesis given by Equation 9 is rejected, then the hypothesis of a treatment effect is supported. If only the hypothesis given by Equation 8 or both hypotheses are rejected, then the data for the experimental group is analyzed as an interrupted time-series experiment.

Discussion

The major purpose of this article was to suggest several improvements to statistical procedures reported by Simonton (1977) for the analysis of time-series designs. The improvements consist of the employment of tests of the fit of proposed models to the data and the *exact* small sample distributional

theory for testing hypotheses. In addition, it should be pointed out that the assumptions underlying the procedure advocated in this article are different from those made by Simonton. Simonton assumed that the errors, $e_1, e_2, \ldots, e_p$, follow a first-order auto-regressive scheme, that is, $e_t = \rho e_{t-1} + \delta_t$, where the residual δ_t is assumed to be independent of other δs and the es. Since N individuals are measured on p occasions ($N > p$), Simonton recommends estimating ρ for each individual (based on p observations) and pooling these estimates across the N individuals. However, when $N > p$, unless there is a strong reason to believe that the particular error structure exists, it may be more meaningful to relax this assumption, assume that the vector of errors $\mathbf{e}$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and dispersion matrix Σ and proceed as suggested in this article. Another advantage of this approach is that the exact distributions of the statistics for testing hypotheses of interest are known (as compared with the Simonton approach in which only the approximate distribution of the test statistic is known).

A necessary condition for the application of the procedures developed in this article is that the number of subjects exceed the number of occasions. Simonton tacitly assumed this condition also. This does not seem to be a serious drawback, since with a large number of subjects, fewer time points are required for satisfactory estimation of parameters. When $N < p$, a structure has to be imposed on $\mathbf{e}$. The procedure suggested by Simonton is more appropriate in this case. However, even in this case, the test of the fit of the model to the data should be explored. This is not a trivial problem but may be solved by adapting methods such as those suggested by Krishnaiah and Murthy (1966) and Rao (1967).

If one of the changed-level models fits the data, and a treatment effect has occurred, then the Swaminathan-Algina (1977) analysis should detect this treatment effect because the changed-level models state that a polynomial fits the pretreatment data, whereas a polynominal does not fit the posttreatment data. Hence, Step 2 of the Swaminathan-Algina analysis should indicate a treatment effect. Are the changed-level models then superfluous? In our opinion, they are not. The Swaminathan-Algina analysis asks whether the pretreatment and posttreatment regression curves on time are the same. An affirmative answer strongly indicates that a treatment effect has not occurred. A negative answer indicates that after the treatment occurred, the regression curve shifted. There may be numerous threats to internal validity that provide reasonable explanations of the shift, and the evidence for a treatment effect may be quite weak. The changed-level models postulate the form of the treatment effect. If the changed-level model fits the data and the changed-level parameter is nonzero, then the class of threats to internal validity that are plausible explanations of the form of the regression curve will probably be narrower. Hence, employing the changed-level models as suggested by Simonton with the analytic procedure advocated in this article would provide an efficient method for the analysis of the interrupted time-series designs.

In the Introduction it was noted that if the residuals are correlated, then the ordinary least squares estimator may not be efficient. Rao (1967) has discussed the situations in which the ordinary least squares estimator is more efficient than the modified generalized least squares estimator. However, strictly speaking, application of his results requires knowledge of the structure of Σ. Furthermore, even when ordinary least squares estimators are more efficient, it can be shown that multivariate procedures should be used for testing hypotheses about the adequacy of the model and about β (Grizzle & Allen, 1969).

References

Algina, J., & Swaminathan, H. A procedure for the analysis of time-series designs. *Journal of Experimental Education*, 1977, *45*, 56–60.

Campbell, D. T. From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

Grizzle, J. E., & Allen, D. M. Analysis of growth and dose response curves. *Biometrics*, 1969, *25*, 357–381.

Jöreskog, K. G. Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology*, 1970, *23*, 121–145.

Krishnaiah, P. R., & Murthy, V. K. Simultaneous tests for trends and serial correlations for Gaussian Markov residuals. *Econometrica*, 1966, *34*, 472–480.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.

Rao, C. R. Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, 1959, *46*, 49–58.

Rao, C. R. Least square theory using an estimated dispersion matrix and its application to measurements of signals. In, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics*. Berkeley: University of California Press, 1967.

Simonton, D. K. Cross sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin*, 1977, *84*, 489–502.

Swaminathan, H., & Algina, J. Analysis of quasi-experimental time-series designs. *Multivariate Behavioral Research*, 1977, *12*, 111–131.

Received February 21, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 927–928

Reply to Algina and Swaminathan

Dean Keith Simonton
University of California, Davis

Algina and Swaminathan have proposed more sophisticated analyses for the cross-sectional time-series experiment. Especially valuable is their suggested procedure for testing the empirical adequacy of the hypothesized intervention model. Nonetheless, the greater complexity of their approach may not always be justified in many research applications. In particular, their exact-test method will normally yield statistical inferences similar to those of my approximate-test procedure.

Algina and Swaminathan's (1979) article constitutes a significant contribution to the recent literature on cross-sectional time-series quasi-experiments. Certainly they have offered some reasonable and more sophisticated alternatives to the analyses I had proposed in an earlier article (Simonton, 1977). Particularly notable, in my view, are the procedures they have developed for testing the adequacy of a model's fit to the data, an issue I neglected. Although it may not always be necessary to run such tests, it is easy to imagine many real situations in which such verification would be required. On the other hand, the general analytical procedure they have outlined for parameter estimation and significance tests is definitely more complicated than that I had proposed. Therefore, it is reasonable to ask what specific advantages accrue from such augmented complexity. Apparently, the chief improvements are two in number. In the first place, the significance tests that they have developed are exact, whereas mine are only approximate. Whether one prefers complex exact tests or simple approximate tests may be somewhat a matter of personal choice, at least given our ignorance regarding the degree of approximation in the approximate tests. Also, the exact tests are only exact when the assumptions are exactly met, and hence the distinction is partially obscured.

Requests for reprints should be sent to Dean Keith Simonton, Department of Psychology, University of California, Davis, California 95616.

However, Algina and Swaminathan raise a second and related critical point about their proposed alternative: My solution to the serial dependencies in the disturbances was to postulate a first-order autoregressive scheme, whereas they estimate a variance–covariance matrix without any a priori structure. As they point out, their procedure requires that the number of subjects exceed the number of observations, but this condition is easily fulfilled. Furthermore, they mention situations in which a first-order autoregressive model may not adequately describe the disturbance process (e.g., when there is measurement error). Nonetheless, I think it is reasonable to ask what the consequences of following my procedure are anyway, no matter what defines the true variance–covariance matrix. Here I believe the differences between the two alternative procedures will usually be small. For example, if the disturbances actually are generated by a second-order autoregressive scheme, the significance tests will be only slightly affected (Miklich, Note 1), whereas the chief loss will fall in the area of estimation efficiency (Engle, 1974). Yet at this point in the history of the behavioral sciences, estimation efficiency (i.e., the variance of our parameter estimates) is probably a low priority concern. Moreover, whenever we are dealing with quasi-experiments entailing only one intervention, the significance tests for cross-sectional time series are extremely robust, even when the autoregression is moderate (Miklich, Note 1). Indeed, it is my belief

that most cross-sectional time-series experiments could probably do without generalized least squares estimation and could simply rely on a more conservative alpha level (e.g., minimum of .01).

In all, I am fairly confident that in the majority of data analysis situations, it may not make a substantial difference whether one employs my simple approximate-test approach with an a priori disturbance model or the Algina-Swaminathan complex exact-test approach with an a posteriori variance–covariance matrix. Nevertheless, I am also of the opinion that their procedures represent a wider range of valuable tests that probably render their approach far more useful in the long run as a general strategy for analyzing such data.

Reference Note

1. Miklich, D. R. *Robustness of analysis of variance treatments: Comparisons to within subjects autoregressive data.* Unpublished manuscript, 1978. (Available from National Asthma Center, 1999 Julian Street, Denver, Colo. 80204).

References

Algina, J., & Swaminathan, H. Alternatives to Simonton's analyses of the interrupted and multiple-group time-series design. *Psychological Bulletin,* 1979, *86,* 919–926.

Engle, R. F. Specification of the disturbance for efficient estimation. *Econometrica,* 1974, *42,* 135–146.

Simonton, D. K. Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin,* 1977, *84,* 489–502.

Received February 20, 1979 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 929–963

The MMPI As a Primary Differentiator and Predictor of Behavior in Prison: A Methodological Critique and Review of the Recent Literature

Milton L. Gearing II
University of South Carolina

Seventy-one investigations of Minnesota Multiphasic Personality Inventory (MMPI) usage in prison work were systematically evaluated. Additional studies were examined to provide a methodological basis for the comparisons of the research, which were made within sections on sampling procedures, sources of variance and their effects on test results, protocol validity, and methods of profile interpretation. Several methodological shortcomings and various differences in procedure across studies limit the generalizability of the findings. However, research in the hostile-assaultive section has produced preliminary MMPI indicators for a type of violently aggressive behavior pattern that is otherwise difficult to detect. Other areas in which the MMPI shows promise include homosexuality, recidivism, and the classification of psychopathologic behavior. More research is needed in the areas of institutional adjustment and suicide. Recommendations for future investigations prescribe adequately controlled sampling procedures, modifications in the interpretation of protocol validity, investigation of certain methodological questions in their own right, consideration of more than one aspect of profile data, the use of base-rate probabilities in predictive studies, and the pursuit of longitudinal studies with thorough follow-up procedures.

The Minnesota Multiphasic Personality Inventory (MMPI) is probably the most widely used personality test in American criminal justice settings today, and MMPI administration is a part of standard admissions practice for all federal institutions as well as for several state and local institutions (Elion & Megargee, 1975). It is routinely used as a general aid in diagnosis and treatment program planning (Haven, Note 1), yet several researchers have investigated more specific applications of the MMPI that deal with several varieties of circumscribed inmate behavioral problems. These investigators are striving to increase the usefulness of the MMPI in prison settings, either for gaining dynamic insights into a specific behavior problem group or for probabilistic prediction of future behavior problems. The full extent of the MMPI's present usefulness and its potential in these two major applications constitutes the focus of the present review. An attempt is made to determine if the MMPI has the potential to become a major and valuable aid in the making of correctional decisions that are appropriate and beneficial for the individual inmate's welfare as well as facilitative of the smooth and effective operation of prison programs in general.

The MMPI consists of 550 different statements covering a wide range of subject mat-

The author gratefully acknowledges the repeated correspondence of James H. Panton and Edwin I. Megargee, whose generous assistance in supplying information undoubtedly improved the quality of this review. The author also acknowledges the helpful correspondence of Paul Gendreau, Thomas E. Hannum, and David M. Pierce.

The author would like to thank Herman Salzberg for his extremely helpful suggestions and Will Drennen and Alexander Galvin for their comments.

Requests for reprints should be sent to Milton L. Gearing II, who is now at the Department of Psychiatry, Division of Psychology, University of Texas Health Science Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75235.

Table 1
Standard Validity and Clinical Scales for the Minnesota Multiphasic Personality Inventory

Symbol	Scale
?	Question
L	Lie
F	Validity
K	Test-Taking Attitude
Hs(1)	Hypochondriasis
D(2)	Depression
Hy(3)	Hysteria
Pd(4)	Psychopathic Deviate
Mf(5)	Masculinity–Femininity
Pa(6)	Paranoia
Pt(7)	Psychasthenia
Sc(8)	Schizophrenia
Ma(9)	Hypomania
Si(0)	Social Introversion

ter. The client responds to each statement by answering *true* or *false* or leaving the statement blank. The standard MMPI profile consists of 4 validity scales (*?, L, F,* and *K*) and 10 clinical scales (*Hs, D, Hy, Pd, Mf, Pa, Pt, Sc, Ma,* and *Si*). The *?* (Question) scale merely records the number of items left blank. The *L* (Lie) scale consists of 15 statements describing minor yet common human failings that attempt to identify intentional efforts by the subject to make himself or herself "look good." The *F* (Validity) scale contains 64 items rarely answered in the scorable direction by normals, and it attempts to detect the presence of confusion due to psychosis or illiteracy, a look-bad attempt meant as a "cry for help" or a random response pattern. The *K* (Test-Taking Attitude) scale consists of 30 items and basically serves as a measure of defensiveness in the subject's test-taking attitude. The *Hs* (Hypochondriasis) scale has 33 items reflecting somatic complaints commonly found in hypochondriasis. The *D* (Depression) scale contains 60 items that describe various symptoms of depression, such as feelings of hopelessness and worthlessness, preoccupation with death, and so forth. The *Hy* (Hysteria) scale is made up of 60 items that tend to identify conversion hysterics in particular. The *Pd* (Psychopathic Deviate) scale is comprised of 50 items designed to detect the amoral and asocial types commonly described as psychopathic personality disorders. The *Mf* (Masculinity–Femininity) scale has 60 items and was initially designed to identify those effeminate males suffering from a sexual inversion disorder but actually seems to reflect aesthetic and vocational interests. The *Pa* (Paranoia) scale contains 40 items intended to detect the clinical pattern of paranoia, which may also be part of another disorder such as schizophrenia. The *Pt* (Psychasthenia) scale is made up of 48 items that attempt to detect the obsessive–compulsive syndrome and that are also suggestive of a high degree of anxiety. The *Sc* (Schizophrenia) scale consists of 78 items that were originally intended to differentiate the psychotic pattern of schizophrenia. The *Ma* (Hypomania) scale contains 46 items that reflect the overactivity, emotional excitement, and flight of ideas common to the affective disorder of hypomania. Finally, the *Si* (Social Introversion) scale is made up of 70 items suggesting unease in social situations, hypersensitivity, and insecurity. The 10 clinical scales are also commonly referred to by number. (See Table 1 for the symbols, numbers, and names of the standard MMPI scales.)

Recently Megargee and his associates (Megargee, 1977a; Megargee, 1977b; Megargee & Bohn, 1977; Megargee & Dorhout, 1977; Meyer & Megargee, 1977) filled an entire issue of *Criminal Justice and Behavior* with the results of 6 years of coordinated research. This research was aimed at the derivation of a comprehensive MMPI classification system for criminal behavior. Megargee (1977a) makes a compelling case for the economy, efficiency, reliability, validity, and operational utility of such a system by comparing it to existing systems on seven dimensions of usefulness that he feels are requirements for productive classification. He then formulates 8 research questions that he asserts must all be answered in the affirmative if the MMPI is to be an adequate basis for a typological system, and he explains how their series of research studies was designed to address these questions. This series of studies has yielded 10 reliably occurring MMPI profile configurations in a population

of youthful male federal offenders. Each profile type is presented with rules for inclusion into its respective classification, modal descriptions of significant characteristics drawn from several assessment sources and case history data, and hypotheses about optimal modes of management and treatment. This computer-assisted system successfully classified 96% of their sample of 1,214 offenders. Although these results are represented as a "progress report," these researchers feel that

> thus far, the taxonomy has surpassed all of our initial expectations. The system has outgrown the capability of one laboratory to investigate it adequately. It is hoped that this series of progress reports will stimulate other researchers to investigate these groups and help us determine their utility. (Megargee, 1977a, p. 113)

Six of the eight questions mentioned above were addressed and have been answered in the affirmative, and this massive research effort continues. Besides seeking replication with other prison populations and in other areas of the country, Megargee and his associates are interested in productively interfacing their fledgling empirical system with both theoretical orientations and other empirical lines of research in corrections (Megargee, 1977b). Their initial results appear promising.

This review focuses on research that has differentiated criterion and comparison inmate groups on some independent basis and has then sought MMPI indicators that reliably differentiate these groups. It is hoped that this will be a useful complement to the work of Megargee and his associates, who are separating inmate groups on the basis of their MMPI profiles and then seeking corresponding dynamic and predictive behavioral indicators. Haven (Note 1) has comprehensively reviewed earlier literature in this area, but a substantial body of research has emerged since his investigation (see Dahlstrom, Welsh, & Dahlstrom, 1975, chap. 3). An up-to-date assessment of the MMPI's potential in prison work is called for, both to stimulate relevant cross-validations and extensions of the latest findings and to maximize the productive interfacing of these findings with the important work of Megargee and his associates. Studies published in English-language journals from 1967 to the time of Megargee et al.'s (Megargee, 1977a; Megargee, 1977b; Megargee & Bohn, 1977; Megargee & Dorhout, 1977; Meyer & Megargee, 1977) publications are emphasized, but studies published before 1967 that meet certain specific criteria (see the last paragraph of the Protocol Validity subsection) are included for the sake of representativeness and continuity. Only studies using some version of the MMPI that at least provides for full scoring of the 13 standard scales are included; Mini-Mult (Kincannon's, 1968, 71-item short form of the MMPI that estimates scores on most of the standard scales) prison research is not addressed. The Review section itself is limited to research that deals with civilian criminal populations incarcerated in conventional correctional facilities or psychiatric facilities specifically for the criminally insane. All criterion and comparison groups are drawn from such populations unless otherwise noted. Research dealing with incarcerated armed services populations, probation populations, and psychopathic/sociopathic populations in conventional mental hospitals is not included.

The first section of this review consists of a methodological evaluation of the research that includes the following subsections: Sampling Procedures, Sources of Variance and Their Effects on Test Results, Protocol Validity, and Methods of Profile Interpretation. Limitations of the research are discussed, and recommendations for standard design modifications and further needed investigations are presented. The studies are then reviewed under two major categories: The MMPI As a Primary Differentiator of Deviant Behavior and The MMPI As a Predictor of Deviant Behavior. A third section, MMPI Differences Across Race and Sex, examines the effects these two variables have on MMPI profiles obtained from prisoners and examines the implications for interpretation and generalization.

Methodological Evaluation

The studies reviewed in this article illustrate several methodological shortcomings.

Table 2
Design Aspects and Subject Data of the Studies Reviewed

Study	V[a]	S[b]	T[c]	Subject data[d]															
Adams (1976)	2[e]	M	a		2			5						11			14		
Adams & West (1976)	2[e]	R	a														14		
Beall & Panton (1956)	1	r	a														14		
Black (1967)	2		d			3		5				9		11			14		
Blackburn (1968)	1[e]		a														14		
Caldwell (1959)	2		s		2	3								11			14		
Carroll & Fuller (1971)	1	R	a		2												14		
Cavior et al. (1967)	1	r	s	1	2		4										14		
Christensen & LeUnes (1974)	1		s														14		
Costello et al. (1973)	1	M	?		2	3		5						11	12		14		
Craddick (1962)	1[e]		s		2	3		5	6					11			14		
Cubitt & Gendreau (1972)	1		s			3											14		
Davis (1971)	1		a														14		
Davis & Sines (1971)	1		a														14		
Deiker (1974)	1[e]	M	s		2	3					8			11		13	14		
Driscoll (1952)	1		a														14		
Dunham (1954)	1		a														14		
Edwards (1963)	1	M	a		2	3		5						11			14		
Elion & Megargee (1975)	3	r	a														14		
Erikson & Roberts (1966)	3		s		2		4	5									14		
Fisher (1969)	1		a					5									14		
Flanagan & Lewis (1974)	2		?														14		
Frank (1971)	2	r	a														14		
Gendreau & Gendreau (1970)	1		a														14		
Gough et al. (1965)	1	r	a														14		
Gregory (1974)	1	r	a														14		
Gynther (1962)	1		a		2			5									14		
Haven (Note 10)	1	R	a														14		
Joesting et al. (1975)	1		a																
Johnson & Cooke (1973)	1		a														14		
Lefkowitz (1966)	1[e]	r	s		2			5									14	15	
McCreary (1975)	1		a		2	3											14		
McCreary & Padilla (1977)	2	R/M	a			3											14		16
Mack (1969)	3		a		2			5		7							14		
Megargee & Cook (1975)	2	r/M	a		2									11			14		
Megargee et al. (1967)	2	r/M	a		2									11			14		
Megargee & Mendelsohn (1962)	2	R	a														14		
Oliver & Mosher (1968)	1		s							7		9		11			14		
Panton (1958)	2	r	a														14		
Panton (1959a)	2		a														14		
Panton (1959b)	2		a														14		
Panton (1960b)	2	M	a		2			5									14		
Panton (1962a)	1	M	a		2	3		5					10				14		
Panton (1962b)	1	m	a		2			5									14		
Panton (1962c)	2		a														14		
Panton (1972)	2	m	a		2			5									14		
Panton (1973)	2	r	a/s														14		
Panton (1974)	2	M	a		2	3		5						11					
Panton (1976)	2	M	a		2	3		5									14		
Panton (1977)	2		a			3		5									14		
Panton (1978)	1	r	a														14		
Panton (Note 6)	1		a														14		
Panton (Note 7)	2		a														14		
Panton & Behre (1973)	1	r	a											11			14		

Table 2 (*continued*)

Study	V[a]	S[b]	T[c]	Subject data[d]						
Panton & Brisson (1971)	2	M	s	2		5		11	14	
Persons & Marks (1971)	1		a						14	
Pierce (1971a)	1	R	a	2					14	
Pierce (1971b)	1	m	$\frac{a}{s}$				7		14	
Pierce (1972b)	1		a	2	3	5			14	
Pierce (1972c)	1	M	a	2	3	5	7		14	
Pierce (1973)	1		$\frac{a}{s}$	2	3	5			14	
Pierce (Note 4)	1		$\frac{a}{s}$	2	3	5			14	
Randolph et al. (1961)	2		s						14	
Rosenblatt & Pritchard (1978)	1	M	$\frac{a}{d}$			5		11	14	
Shinohara & Jenkins (1967)	3[e]		s						14	
Snortum et al. (1970)	1		a						14	
Stanton (1956)	2[e]		a					11	14	16
Stump & Gilbert (1972)	1		a						14	
Sutker & Moan (1973)	1	r	s						14	
Tsubouchi & Jenkins (1969)	3		s						14	
Twomey & Hendry (1969)	2	R	s						14	
Wattron (1963)	1	r	a						14	
Wilcock (1964)	1	M	?				7		14	

[a] V = validity criteria: 1 = no criteria employed; 2 = employed standard cutoffs (*L* scale *T* score maximum of 70, *F* scale *T* score maximum of 80, *K* scale *T* score maximum of 70) or discarded invalid Minnesota Multiphasic Personality Inventories (MMPIs); 3 = detected and discarded random profiles only.
[b] S = sampling procedures: R = complete randomization; M = complete matching; r = partial randomization; m = partial matching; $\frac{r}{M}$ = mixed approach; more than one pair of criterion and comparison groups used.
[c] T = timing of MMPI administration: a = routine admissions administration; s = given specifically for the study conducted; d = administered just prior to discharge; ? = not specified; $\frac{a}{s}$ = mixed approach, more than one pair of criterion and comparison groups used.
[d] Those characteristics either controlled or matched: 1 = addiction status; 2 = age; 3 = educational achievement; 4 = ethnicity; 5 = IQ; 6 = length of current sentence; 7 = length of time in prison; 8 = marital status; 9 = number of prior convictions; 10 = number of disciplinary restrictions; 11 = race; 12 = rate of recidivism; 13 = religion; 14 = sex; 15 = social class; 16 = socioeconomic status.
[e] Conducted preliminary IQ/reading-level screening.

In addition to the problems discussed later, the studies vary on such dimensions as timing of administration of the MMPI (i.e., on admission vs. at the time of the research study), security grade of the institution from which the test groups were drawn (this was frequently not specified) and region of the country in which the study was conducted. Table 2 presents most of the important characteristics of the studies reviewed, and the lack of consistency with respect to the control of subject characteristics across the studies is readily apparent. Overall, the variations in procedure substantially limit the generalizability of present results.

Sampling Procedures

Slightly more than half of the studies used some form of matching (i.e., matching on a few dimensions such as age, length of imprisonment, IQ, or educational level) or random sampling, whereas the rest used every available member of a group with completed MMPIs. Those studies employing partial randomization frequently used ran-

domly selected comparison groups while using all available prisoners who qualified for the criterion group. Studies employing partial matching commonly had more than two groups within an experimental comparison or more than one comparison and only matched some of their groups on a few dimensions, such as those mentioned above. (See Table 2 for all of the dimensions that were employed.)

Although complete randomization in prison research is sometimes impractical, partial approaches such as those above do not compensate for systematic bias. If full randomization cannot be realized, then matching a comparison group to the available criterion group (which is defined by a history of a given target behavior such as escapism, homosexuality, etc.) is preferable to randomizing the comparison group, since the latter procedure does not constitute an intelligible comparison. Partial matching approaches have often been necessitated by the lack of data on some matched variables for certain groups, yet this still serves to weaken the representativeness of results. In short, the findings of many studies are severely limited in their generalizability and must be interpreted with caution.

Sources of Variance and Their Effects on Test Results

Table 2 demonstrates that the only source of variance that was consistently controlled across almost all studies was sex. Sixty-nine studies limited their observations to only one sex, and the overwhelming majority investigated male prisoners. The only other sources of variance that were controlled by a substantial number of studies were age, educational achievement, IQ, and race.

Some research has been done on the effects of these latter variables on MMPI profiles in general. Costello, Tiffany, and Grier (1972) examined several methodological questions with respect to racial (i.e., black–white) comparisons in their investigation of carefully matched inpatient and outpatient psychiatric clients. They found that blacks had more elevated profiles in general, with significantly higher scores on the *F*, *Hs*, *Mf*, *Sc*, and *Ma* scales and significantly lower scores on the *K* scale. Gynther (1972) reviewed the literature on black–white MMPI differences in normal populations and reached several pertinent conclusions. There seem to be consistent differences between blacks' and whites' MMPIs, the most frequent being significantly higher elevations for blacks on the *Sc* and *Ma* scales, which seem to be affected by such variables as education, residence, and cultural separation. However, Gynther claims that there is no evidence to indicate that these trends signify that blacks are less well adjusted than whites. Gynther advocates the construction of a new MMPI form with *T* scores based on black norms and with the derivation of black behavioral correlates based on profile types. He suggests that a temporary solution would be the generation of qualifying rules for the interpretation of blacks' MMPIs, and he urges that those interpreting MMPIs should be alerted to the potential misuse of the test in treating black profiles in the same manner as white profiles.

Subsequent research has attempted to clarify the issues raised by Gynther's (1972) review. Gynther, Altman, and Warbin (1973) investigated the consistent finding of higher *F* scale scores for blacks by comparing the MMPIs of 1,125 white psychiatric inpatients to those of 134 black psychiatric inpatients (63% of the overall sample was male). They also used records that contained sets of 168 demographic descriptors and 111 mental status items for each patient. Nonrandom profiles having an *F* scale raw score $\geqslant 26$ were separated from the other profiles for each racial group, and the demographic and mental status items were analyzed for their ability to differentiate the groups. Replications were carried out for both black and white groups. For whites, 16 items overall provided significant differentiations across the original and replication samples, and these items essentially characterized *F* scale $\geqslant 26$ profiles as representing "confused psychotics." This was consistent with earlier research findings. However, the original black sample with *F* scale $\geqslant 26$ profiles shared only 5 discriminating items with the original white sample, and only 3 of these items were scored

in the same direction. Furthermore, not one variable provided a significant differentiation for both the original and replication samples of blacks. Gynther et al. (1973) concluded that blacks with an *F* scale score $\geqslant 26$ are not seen any differently than blacks with lower *F* scale scores when compared on a mental status exam, and they maintain that an *F* scale score $\geqslant 26$ for a black psychiatric patient has different behavioral correlates than a similar score for a white psychiatric patient. They also suggest that this difference may be due to "race-sensitive" items on the MMPI; blacks seemed to favor items reflecting certain attitudinal dispositions such as alienation, religiousness, liking for school, and romanticism.

As part of a larger project, Gynther, Lachar, and Dahlstrom (1978) used a large sample of normal, conservative middle-class black adults from Alabama, Michigan, and North Carolina (321 males, 561 females) to generate a new *F* scale for blacks. They replicated the criterion used when the original *F* scale was generated on a Minnesota sample, which accepted items answered in a given direction by 10% or less of the sample. They identified a total of 33 items in this manner, 22 of which are on the standard 64-item *F* scale. They found that higher scores in their black sample for both the original and the new *F* scale were significantly related to a younger age, less education, and a less-skilled job classification of the head of the household. Gynther et al. (1978) concluded that the significant correlates of high *F* scale scores for whites do not appear to be valid for blacks, and they suggest that their new *F* scale may be a better measure of deviant responding for blacks than the standard *F* scale.

Further studies have investigated other MMPI scales as well in their examination of racial differences on the test. Penk and Robinowitz (1974) investigated the profiles of black and white male veterans who were drug abusers. They compared black opiate users to white opiate users as well as comparing black opiate users to white opiate nonusers. They found that black opiate users scored significantly lower than white opiate users on the *F* and *Hy* scales and scored lower than white non-opiate users on the *F*, *Pt*, *Sc*, and *Si* scales. They note that these findings conflict with the expectation that blacks typically obtain higher MMPI scale scores.

Davis and Jones (1974) compared black and white veterans who were psychiatric patients by systematically varying race, education, and differential diagnoses obtained independently from the MMPIs. An initial survey of all available black subjects (390) and of a randomly selected white comparison group found that blacks were more likely to be diagnosed schizophrenic, whereas whites were more likely to be diagnosed alcoholic and/or depressed. They concluded that there seem to be race-related differences in diagnosis, but they were unable to determine whether this was due to actual race-related variations in psychopathology or to blacks being misdiagnosed more often. Next, Davis and Jones randomly selected 20 subjects for each of eight groups; each racial group was blocked into groups of high-education (minimum of 12 years of education) schizophrenics, low-education schizophrenics, high-education nonschizophrenics, and low-education nonschizophrenics. Subjects older than 50 or having an *F-K* index greater than 14 were not included. They found that there was no significant main effects associated with differences in race. Schizophrenics scored significantly higher on the *Pa* and *Sc* scales than nonschizophrenics, and the poorly educated scored significantly higher on the *Sc* scale than the well educated. Some interaction effects were evident: The more poorly educated blacks and better educated whites scored significantly higher on the *Pa* scale than the other subjects, and the more poorly educated blacks scored significantly higher on the *Sc* scale than the other subjects (with the better educated blacks scoring the lowest on this scale). Davis and Jones concluded that race is a type of independent subject variable that confounds attempts to adequately interpret MMPI profiles of randomly selected black versus white groups, but this is *not* the case when diagnosis and education are controlled. They maintained that with poorly educated groups or with randomly selected groups in which the majority of blacks are poorly educated, blacks would be

expected to perform in a more pathological direction than most random samples of whites who are similar in characteristics, such as type of psychopathology.

Cowan, Watkins, and Davis (1975) examined the same subjects used by Davis and Jones (1974) and blindly sorted the MMPI profiles into schizophrenic and nonschizophrenic groups. A protocol was classified as schizophrenic if the *Sc* scale *T* score exceeded 70 and if the *Sc* scale was more elevated than the *Pt* scale. In three of the four nonschizophrenic groups, the profiles were correctly classified significantly more often than by chance in a way that conformed to previously determined expectations for percentage of correct classification. However, almost one half (9 out of 20) of the poorly educated black nonschizophrenics were misclassified as schizophrenic. Cowan et al. concluded that although cultural background factors exert a significant influence, much of the reported variation in blacks' MMPIs may be a function of education. They maintained that the MMPI appears to retain adequate discriminative power for blacks with at least 12 years of education.

The conclusions of Davis and Jones (1974) and Cowan et al. (1975) seem to constitute the key issue here. Davis and Jones suggest that their results indirectly support Gynther's (1972) hypothesis, which maintained that blacks produce more apparently pathological MMPIs because of their experiences of alienation from the established white society. Gynther maintains that blacks learn to be suspicious of others because it is necessary for their survival in this society. However, Davis and Jones point out that well-educated blacks do not appear to demonstrate the consistent racial differences on MMPI scales that randomly selected groups of blacks do, and they conclude that a more advanced education seems to have either a masking or obliterating effect on cultural factors that can produce misleading MMPI configurations. They hypothesized that a selection process may be operating here in which the more sensitive and suspicious blacks drop out of school as soon as they are legally allowed to. On the other hand, blacks that elect to continue their education are exposed to extended culturation effects through their prolonged contact with white values and expectations, and since they are less sensitive and suspicious than those blacks who drop out, they are more apt to assimilate the cultural values of whites. Cowan et al. essentially concur with this perspective. Therefore, if higher elevations for blacks on MMPI scales are chiefly due to factors such as differences in IQ or education instead of race, then Gynther's advocacy of separate norms for blacks seems inappropriate. Apparent bias could be due to a combination of factors in which race may play a minor role. At this point the assumption of racial bias of the MMPI appears to be premature at best; the entire issue demands further research. (See the discussion of racial differences in MMPI profiles of prisoners in particular, which is included near the end of the Review section.)

Panton (1960a) has investigated the effects of intelligence on profiles taken from a prison population. Prisoners with an IQ below 110 consistently demonstrated neuroticism and anxiety (elevations on the *Hs*, *D*, *Pt*, and *Sc* scales), whereas prisoners above the 110 cutoff consistently demonstrated character disorders (elevations on the *Hy*, *Pd*, and *Ma* scales). Panton (1959a) also examined the effects of age on prison profiles and found that white inmates over 30 years old scored significantly higher than under-30 white inmates on the *Hs*, *D*, *Hy*, and *Pd* scales and significantly lower on the *Sc*, *Ma*, and *Si* scales. (Groups of black inmates did not show these differences.) A second study by Panton (1976–1977) on the effects of age found that when compared to baseline MMPI figures for the prison population, inmates who were at least 60 years old displayed "a neurotic overlay with less psychopathy" (significantly higher on the *Hs*, *D*, *Hy*, and *Si* scales and significantly lower on the *Pd* scale). In their investigation of the *F* scale in white psychiatric patients, Gynther and Shimkunas (1965a; see the section on Protocol Validity) found an interaction between age and intelligence that affected *F* scale scores.

These findings are far from conclusive, but they do suggest that all of these factors are

primary sources of variance that should be controlled in MMPI research. In fact, they demand further investigation in their own right (especially the issue of racial and cultural bias), but until the nature of their influence is clarified, they should be dealt with either through matching or blocking procedures. Even researchers using full randomized assignment would be well-advised to examine their samples post hoc for significant differences on age, IQ, and race. Other variables in Table 2 that are more specific to prison populations have not been investigated for their effects on MMPI protocols. Therefore, randomization seems to be the preferred method to control for confounding effects, since these latter potential sources of variance may exert a presently unknown influence on test results.

Protocol Validity

Hathaway and McKinley (1967) state that "subjects sixteen years of age or older with at least six years of successful schooling can be expected to complete the MMPI without difficulty" (p. 9). Yet only eight of the studies reviewed (see Table 2) mentioned any attempts to screen their subjects for IQ or reading level, even though there is a significant incidence of illiterate and foreign-language-speaking inmates in prison in general. Such preliminary screening is an important consideration in obtaining valid profiles and should be a standard part of MMPI research in prisons. Dahlstrom, Welsh, and Dahlstrom (1972) recommend the use of brief intellective or reading achievement measures such as the Wide Range Achievement Test (Jastak, Bijou, & Jastak, 1965), the Ohio Literacy Test (Foster & Goddard, 1924), or the Kent EGY (Kent, 1946) to identify subjects who may have difficulty completing the MMPI (p. 21). However, they report one study (Glenn, 1949) that investigated the use of oral presentation of the MMPI to some retarded juvenile delinquents who lacked the above minimal educational requirements. This study concluded that success could be achieved for subjects with an IQ of at least 65 and with at least 3 years of schooling.

Dahlstrom et al. (1972) also report that Panton has adopted an oral delivery approach for the group testing of prisoners in which the inmate with the most education and best reading competency reads the items to the rest of the group. The extent to which this procedure is legitimate with respect to those prisoners below the IQ/reading-level criteria has not been investigated. Therefore, at this time the use of the MMPI with prisoners failing to meet these criteria cannot be sanctioned as a sound methodological practice. Further research should investigate the advisability of oral administration with prisoners having less than a sixth-grade reading level or an IQ below 80 (the common Wechsler Adult Intelligence Scale [WAIS] cutoff; see Dahlstrom et al., 1972, p. 21). One alternative for researchers whose sample representativeness is threatened by this requirement would be to administer oral versions of the MMPI to the subjects in question and analyze their overall outcome data twice, once by including these subjects and once by excluding them from the analysis. Differences in results could then be examined, and unless the number of such oral administrations is substantial, few if any significant differences in outcome between the two analyses should be found.

One study links the validity of a given MMPI profile with the timing of its administration in the prison setting. Pierce (1972a) gave 60 inmates the MMPI on their admission to prison and then retested the same group 6 weeks later. He found that group mean scores dropped significantly on the *Pt*, *Sc*, and *Si* scales, whereas the group mean score rose significantly on the *K* scale. He concluded that the initial profiles were confounded by the presence of stress associated with entering prison, and this tended to obscure the actual profile configuration that was indicative of the inmate's true personality. Although comparisons of changes within an individual inmate's two profiles would have been preferable to Pierce's method of comparing differences between the group means for each scale, Pierce has raised an important question here. Further research should determine whether the delaying of MMPI administration to allow

for some institutional adjustment and a corresponding decrease of situational stress would result in more accurate and more useful MMPI profiles.

The *F* scale is probably the single most widely used indicator of MMPI profile invalidity. A common contemporary practice is to discard profiles with *T* scores exceeding 70 for the *L* scale, 80 for the *F* scale, or 70 for the *K* scale, and a high *F* scale score is usually the criterion that causes the discarding of protocols. As shown by Table 2, 29 of the previously mentioned studies deal with this protocol validity question in some manner, whereas the rest seem to ignore the issue entirely. However, research suggests that the rigid application of the *F* scale *T* score exceeding 80 to determine invalidity may be ill-advised. Comrey's (1957–1958) factor analysis of the *F* scale led him to conclude that a high *F* scale score may be a valid indicator of pathology, not a signal of profile invalidity. Morrice (1957) felt that high *F* scale scores were an index of personality disorder in his group of recidivist criminals, and he suggested focusing on overall profile configurations as meaningful while underinterpreting absolute single scale elevations. He reports the following "impressionistic" investigative results:

> The test was repeated in three fully investigated prisoners to determine whether their profiles were reproducible. In fact the test, repeated after an interval of six to seven months, reproduced the original profiles very closely and in each case with a repeat of an abnormally high *F* score. There was no reason for these three prisoners to malinger and it would be clever deception indeed to reproduce nearly identical responses . . . . The impression gained is that a high *F*, together with several abnormal scores on the personality scales, is meaningful in terms of personality disorder of antisocial type. (Morrice, 1957, p. 634)

This latter study was the first suggestion that people specifically classified as character disorders may validly produce high *F* scale scores when they respond to the test in a candid and truthful manner. In their consideration of the meaning of the *F* scale, Dahlstrom et al. (1972) concluded that in certain instances, "elevated *F* scale scores are part and parcel of the behavioral disorder generating the clinical-scale configuration, documenting its range and severity but not reflecting adversely upon the dependability of the MMPI protocol itself" (p. 161; see this source for a more in-depth consideration of the *F* scale). They even suggest that valid profiles may, in "rare instances," exhibit raw-score *F* scale values of 40 or more (*T* score well over 110) for these very reasons.

Later studies supporting the validity of nonrandom profiles with *F* scale *T* scores exceeding 80 have searched for possible clinical interpretations of the high *F* scale score. Gynther (1961) found evidence which suggested that *F* scale scores could differentiate between diagnostic classifications, since his group of "aggressive" criminals tended to get *T* scores exceeding 80 significantly more often than "passive" criminals. Investigating *F* scale scores with respect to 12 crime classification groups, Gynther (1962) also found a significant relationship between high *F* scale scores and sexual crimes and concluded that larger *F* scale values indicate "emotionally 'sicker'" criminals. Blumberg (1967) found supporting evidence for Gynther's (1962) conclusion. However, Gauron, Stevenson, and Englehart (1962) failed to differentiate behavioral disorders from the rest of their hospital patient sample on the basis of high *F* scale scores and concluded that an *F* scale *T* score exceeding 80 "cannot be routinely employed as a diagnostic sign of behavior disorder with psychiatric patients" (p. 488). Gynther and Shimkunas (1965b) attempted to confirm these findings by using both hospital patient and prisoner groups. Their psychotic patients accounted for almost 70% of their *F* scale *T* scores that exceeded 80 for the patient group, whereas their behavior disorder prisoners accounted for 66% of their *F* scale *T* scores that exceeded 80 for the prisoner group. The reversal of this relationship between high *F* scale scores and diagnosis was significant at the .001 level of confidence. They concluded that the differences between the two samples reflect personality features of those who break the law, and they concur with Leary's (cited in Gynther & Shimkunas, 1965b) assertion that the *F* scale measures hostility and aggression. Rice (1968) also found a

significant relationship between *F* scale *T* scores that exceeded 80 and overtly hostile, aggressive behavior.

A few studies have attempted to derive diagnostic indicators from the *F* scale. McKegney (1965) found that 21 specific *F* scale items were answered in the scorable direction by male juvenile delinquents significantly more frequently than by normals. He concluded that these items accurately described special problem areas of the subjects in his sample even though they caused *F* scale *T* scores above 80, and he suggested that specific *F* scale item endorsements could provide clinical insight into individual cases. Dahlstrom et al. (1972) echo this suggestion, pointing out that 21 *F* scale items appear on Grayson's Critical Items List (Grayson, Note 2) and stating that "in the utilization of the *F* scale items as a set of rare answers, the clinician should not lose sight of the possibility that one or all of the answers to these items from some test subjects may be quite literally true, clinically relevant, and worthy of special investigation" (p. 115). Gynther and Petzel (1967) failed to confirm their hypothesis that psychotics and individuals with behavioral disorders with *F* scale *T* scores exceeding 80 endorse different patterns of *F* scale items; only one item (No. 215, "I have used alcohol excessively") reliably differentiated the two groups. However, using this one item in conjunction with the total number of items endorsed on a 12-item *F* scale subscale of manifest psychotic content proved to effectively differentiate the two groups. They suggest that a "nonconformity" dimension causes high *F* scale scores and is a general dimension underlying both psychosis and behavior disorders.

The effects of age and intelligence on *F* scale scores in psychiatric patients were investigated by Gynther and Shimkunas (1965a). Several of the above studies have found that youth correlated with higher *F* scale scores (Blumberg, 1967; Gauron et al., 1962; Gynther, 1961; Gynther, 1962), but no consistent relationship was found between high *F* scale scores and intelligence (Gynther, 1961; Gynther, 1962) or educational level (Gauron et al., 1962; McKegney, 1965). Gynther and Shimkunas (1965a) found an interaction between age and intelligence that they termed the "critical element" affecting changes in *F* scale scores: The scores decreased with increasing age for low- and high-IQ subjects but remained relatively constant for average-IQ subjects. However, educational level did not effect *F* scale scores.

A few studies have examined other indices of invalid MMPI profiles with prisoners. Lawton and Kleban (1965) retested their prisoner group with the MMPI and told them to simulate someone who had never been in trouble with the law. Seven clinical scales showed significant drops compared to the original testing, but relative configurational elevations did not change significantly. Lawton and Kleban concluded that prisoners are unable to successfully manipulate the *Pd* scale alone to conceal their sociopathy. Bennett (1970) used a similar design and concluded that inmates either cannot or do not "fake good" to any significant degree. However, Gendreau, Irvine, and Knight (1973) criticized the above instructional sets as unrealistic and after obtaining protocols in the standard fashion, instructed their prisoners to successively fake bad and fake good on retests as if they were trying to manipulate the prison system for desired treatment or privileges. They found that faking in both directions radically distorted initial profiles and successfully concealed the basic characterological problems of the prisoners. Both the *F* scale (using a raw score cutoff of 34, which has a *T* score well over 100) and the *F-K* index (subtracting the *K* scale score from the *F* scale raw score and using a cutoff of 24) successfully separated 100% of the fake-bad profiles from the original standard administration profiles. All seven of the indicators that they examined effectively identified fake-good profiles, the two best indicators being the Positive Malingering (*Mp*) scale (92% overall hit rate) of Cofer, Chance, and Judson (cited in Dahlstrom et al., 1972) and the *F-K* index (85% hit rate). Ability to fake effectively was not found to be related to IQ. Gendreau et al. advocated further research investigating the readjustment of the standard *F* scale and

F-K index cutoff scores to properly discriminate honest versus faked inmate profiles and suggested the possibility of the routine use of the *Mp* scale in identifying fake-good profiles.

As part of their larger study investigating racial differences on the MMPI, Costello et al. (1972) considered the effects that different validity criteria for the selection of protocols have on research outcomes. Their study analyzed the data by using all available protocols and then reanalyzed the data by using only valid protocols (*?* scale maximum *T* score of 100, *L* scale maximum *T* score of 70, *F* scale maximum *T* score of 80, *K* scale maximum *T* score of 70, *F-K* maximum raw score of 9). They found that all significant differences in the initial analysis were eliminated in the reanalysis, and they felt that this was partially due to the fact that high *F* scale scores are associated with elevations on certain clinical scales (such as the *Sc* scale). They concluded that the employment of validity criteria similar to those described above restricted both profile variability and the detection of actual differences between groups. (They also noted that more black profiles were eliminated by those criteria than white profiles.)

The above studies offer several insights into what constitutes a valid MMPI inmate profile. Of primary importance is the finding that the conventional cutoff of the *F* scale *T* score exceeding 80 should not be dogmatically employed as a criterion for profile invalidity; profiles with this *F* scale elevation may actually be valid, especially with a prison population. A high *F* scale score does not seem to necessarily indicate any single behavior or diagnosis but may be associated with a generally nonconforming, hostile, and aggressive approach to life. Although examination of *F* scale item endorsements on individual profiles may provide useful clinical insights, the use of high *F* scale scores in differential diagnosis does not seem to be a promising possibility. Factors such as age and intelligence may also effect *F* scale elevations, but more research using prison populations specifically is needed, especially to check for any interaction effect resembling that found by Gynther and Shimkunas (1965a). Dependable indicators of faked prisoner profiles are sorely needed, but none have been identified to date.

Megargee (Note 3) described a complex approach to the profile validity problem that he and his associates have used in the development of their classification system. Reading levels for prisoners were not determined through any routine testing methods, but those inmates who were found to have reading difficulties were tested orally with a tape-recorded version of the MMPI. The Spanish-speaking inmates were given the Spanish version of the MMPI. After each inmate completed the MMPI, he was required to correctly identify his answers to six test items randomly chosen from his protocol by the examiner. Failure to pass this check necessitated a retest. Profiles with *F* scale *T* scores exceeding 100 were examined clinically for their approximation to the mean random-response clinical-scale pattern, which consists of scores on each scale that are equivalent to the midpoint (i.e., half the items marked in the scorable direction) of that scale. Only those profiles that deviated from this pattern on the validity scales and the clinical scales and that made "clinical sense" were retained as nonrandom.

This approach was the best one encountered by this author, but it still has some drawbacks. For one thing, Megargee states that "obviously some expertise with the MMPI is involved" in the latter clinical scrutinizations of suspect profiles; the validity sections of Dahlstrom et al. (1972) may aid the researcher who lacks such expertise in coming to grips with this challenging interpretive task. Contrary to the previously mentioned IQ/reading-level recommendations, no intelligence testing was done to screen inmates for the MMPI. Again, the testing of subjects lacking the minimal reading-skills criteria has not yet been empirically demonstrated to be a sound methodological practice, and the lack of preliminary reading-level screening cannot be sanctioned. Even so, this approach is more acceptable overall than most being employed today, and the combination of routine screening examinations with Megargee's guidelines is probably the best approach to the issue of MMPI protocol

validity that can be realized presently. One alternative that may prove to be a more economical means for the detection of randomly answered profiles (especially if protocols are being drawn from an accumulated data bank) is the test–retest (TR) index of Buechley and Ball (1952). This index examines the 16 repeated items in the booklet and R forms of the MMPI for consistency of item endorsement and employs a cutoff of more than three disagreements to identify and discard randomly generated profiles. Unfortunately, the lack of dependable fake-bad and fake-good indices cannot be satisfactorily compensated for at this time; the findings of Gendreau et al. (1973) must be cross-validated and built upon to fill this void. Future studies should also investigate the retesting approach for random profiles mentioned above and should consider the use of standardized instructions cautioning against any further attempts at deception. The comparability of such protocols obtained under added duress with protocols obtained in the standard manner would be difficult to assess, however. (See Dahlstrom et al., 1972, pp. 105–106, for a discussion concerning the implications of the forced-choice change in administration; for a more thorough treatment of all of the above validity indicators as well as other validity indicators that have not yet been investigated in the prison population, see the validity sections in Dahlstrom et al., 1972.)

In summary, there is a strong possibility that a good deal of the MMPI research with prisoners has been adversely effected either by the failure to employ any criteria for profile invalidity (causing a confounding of any actual differences between groups by random profiles) or by the overly rigid application of conventional validity criteria (causing a concealment of significant differences between groups by restricting elevation ranges on several scales). The validity criteria used in each study reviewed in the main body of this article will be indicated by a number after the date of each article the first time it is cited (e.g., Beall & Panton, 1956; 1). This number corresponds to the number entered under the V column of Table 2: 1 = no criteria employed; 2 = employed standard cutoffs (*L* scale *T* score maximum of 70, *F* scale *T* score maximum of 80, *K* scale *T* score maximum of 70); 3 = detected and discarded random profiles only. This is intended to aid in the consideration of the possible effects that selection of validity criteria may have had on the findings presented. The present author advocates Alternative 3 as the most appropriate procedure until further research on fake-bad and fake-good indices clarifies their proper application with inmate MMPI profiles.

As mentioned in the introduction, special validity criteria were used in the selection of pre-1967 studies for review. Any such study that either employed standard cutoffs (usually, *L* scale maximum *T* score of 70, *F* scale maximum *T* score of 80, and *K* scale maximum *T* score of 70) or at least provided validity scale data (usually means and standard deviations) were included, since this allows at least a rough assessment of the direction and extent to which the findings may be distorted. Pre-1967 studies that failed to meet these criteria provide no clues for such an assessment and were therefore omitted to keep the present review to a manageable length; the only exceptions made were for studies that generated original experimental scales, for these studies were necessary for purposes of continuity in the review. (See Haven, Note 1, for a more comprehensive review of pre-1967 studies.)

Methods of Profile Interpretation

One crucial choice that any MMPI researcher must make is which of the several ways of viewing MMPI data will maximize the quality and quantity of useful information gained from the protocols. The studies reviewed in this article consider one or more of several protocol aspects, including conventional scale elevations, mean profile configurations for each group, actuarial high-point coding systems, a sequential linear-sums model, and experimental scales using cutoff scores. Costello et al.'s (1972) previously reviewed study on racial differences (see the section on Sources of Variance and Their Effects on Test Results) examined the variations within the significant findings of

several different data assessment approaches that were applied to the same set of MMPI protocols. They found that consideration of isolated scale elevation means between groups produced significance on the *F, K, Hs, Mf, Sc,* and *Ma* scales, whereas a high-point coding approach identified the *Mf, Pa, Sc,* and *Si* scales, and a two-digit coding analysis identified the *Hs, D, Pd, Pa, Pt,* and *Sc* scales. They concluded that "inferences drawn from contrasting differences would depend on the particular dependent measure employed" (p. 167). Black's (1967) study (see the section on Recidivism) serves as a valuable object lesson in this respect; even though initial investigation of isolated scale elevations failed to produce significant differences between his recidivist and nonrecidivist groups, Black carried out further analyses in a highly resourceful manner until he obtained a combination of indices that correctly identified 90% of his overall sample. Efficient maximization of profile data may demand the investigation of several different approaches simultaneously, for the above two studies suggest that not only the nature of interpretation of results but even the efficient detection of existing significant differences are a function of the type of MMPI dependent measure employed.

Gregory's (1974; see the section on Classification of Psychopathologic Behavior) unique sequential linear-sums approach also deserves consideration here. Essentially, this approach used a stepwise regression analysis that employed the conventional MMPI scales as predictor variables and the classifications based on checklist ratings of available case history data as criterion variables. His three index formulas successfully classified 63% of his overall sample as psychopathic, adjusted, or neurotic. Gregory's perspective is that

> the evaluation of code type systems should be pragmatic, that is, based on the proportion of target population profiles that can be interpreted within the system and on the degree to which stable and useful personality correlates are generated . . . *utility* must always have the final say. (p. 391, italics in original)

The extent to which Gregory's approach lives up to this standard as compared with other approaches needs to be examined with further comparison research in the prison population.

In their important and extensive work on the efficiency of predictive psychometric indicators, Meehl and Rosen (1955) identified a critical methodological point, not observed by most of the predictive studies in this review, that used "cutting scores" on experimental scales. Such studies usually isolate those MMPI items that differentiate best between their research groups and use these items to form an experimental scale. The distribution of total scores on the scale is examined for each of the two groups, and the single score that maximizes the differentiation of the two groups (i.e., the first group mostly falls on one side of this score, whereas the second group mostly falls on the opposite side of this score) is designated the "cutting score" for the scale. However, Meehl and Rosen point out the necessity of considering the base rate of occurrence of a given criterion variable in the overall population under study in the investigations of this type. This is required because a psychometric predictor such as an experimental MMPI scale that at first blush appears to make an impressive differentiation may actually cause more incorrect identifications than the base-rate differentiation. This is particularly a problem when the criterion variable normally occurs in almost all or almost none of the population under study, whereas it ceases to be a problem when the criterion variable normally occurs in approximately 50% of this population.

Shupe and Bramwell (1963; this study did not meet the validity criteria and is not included in the Review section) constitute an example of this problem in their investigation of the Prison Escape (*Ec*) scale. Since the base rate of escape at their institution was about 5%, their sample results indicated that the *Ec* scale could predict escape risk in the overall population with 90% accuracy. However, Shupe and Bramwell realized that even though they could correctly predict escape risk in all prisoners 90% of the time by using the *Ec* scale, they would correctly predict escape risk in all prisoners 95% of the time if they predicted

that no inmates would escape. Another problem with the *Ec* scale is that it would produce about 7.6% false positives (nonescapees incorrectly labeled as escapees) in the overall population, whereas base-rate prediction would produce no false positives, since it predicts that no one will escape. These findings are a powerful illustration of Meehl and Rosen's disquieting assertion that "*deciding on the basis of more information can actually worsen the chances of a correct decision*" (p. 202, italics theirs). In view of this methodological revelation, Gough, Wenk, and Rozynko's (1965; see the section on Recidivism) prescription seems most appropriate:

> All claims as to predictive accuracy must be stringently verified by utility analyses. In essence, this means that in any prediction a chance level of accuracy must be defined, based on the observed frequency of the criterion, and then the diagnostic or forecasting technique must be contrasted with this chance level. (p. 433)

Any researcher involved in predicting anything from MMPI inmate profiles should incorporate this crucial point and Meehl and Rosen's other methodological requirements and helpful suggestions into their experimental design (see also Cronbach & Gleser, 1965). Unfortunately, Gough et al.'s study is the only work of a predictive nature reviewed in this article that followed Meehl and Rosen's recommendations. The other predictive researchers reviewed here should obtain appropriate base-rate figures from their overall sample population and reevaluate their findings following Meehl and Rosen's prescriptions. The true utility of their findings could not be assessed adequately in every case in this article, since few studies provided the appropriate base-rate figures. Although Meehl and Rosen point out that certain special situations (e.g., see the discussion on suicide in the section on Institutional Adjustment) mitigate the relevant applicability of chance level prediction accuracy, their guidelines should be considered nonetheless; the pros and cons of using base-rate prediction in any situation must be individually assessed for the particular situation in question.

Table 3
Experimental Scales Among the Significant Results of the Studies Reviewed

Symbol	Scale
A	Anxiety
AI	Anxiety Index
Al	Alcoholism
Ap	Prison Adjustment
As	Asocial
Asx	Aggravated Sex
At	Anxiety
CI	Critical Item
CR	Conversion Reaction
DaS	Drug Abuser
Dc&i	Defect of Inhibition Control
DH	Direction of Hostility
Dn	Denial
Ec	Prison Escape
Em	Emotional Immaturity
Eo	Ego Overcontrol
ES	Ego Strength
Ex	Extraversion
FTI	Frustration Tolerance Index
GH	General Hostility
HC	Habitual Criminal
He	Heroin
Hsx	Homosexual
Hy2	Need for Affection and Reinforcement subscale
Hy3	Lassitude-Malaise subscale
In	Inner Maladjustment
Jh	Judged Manifest Hostility
Ma1	Amorality subscale
Ma2	Psychomotor Acceleration subscale
Ma3	Imperturbability subscale
Ma4	Ego Inflation subscale
Mf1	Personal and Emotional Sensitivity subscale
Mf5	Denial of Masculine Occupations subscale
Mp	Positive Malingering
O-H	Overcontrolled Hostility
Pa1	Ideas of External Influence subscale
Pa2	Poignancy subscale
PAS	Prison Adjustment
PaV	Parole Violator
Pd1	Family Discord subscale
Pd2	Authority Conflict subscale
Pd4A	Social Alienation subscale
PM	Prison Maladjustment
Pq	Psychotic Tendency
Pr	Prejudice
R	Repression
Re	Responsibility
Rmn	Recidivism-Rehabilitation
R-S	Repression-Sensitization
Sc1A	Social Alienation subscale
Sc2A	Lack of Ego Mastery-Cognitive subscale
SD	Sensorimotor Dissociation

Table 3 lists all of the abbreviations and corresponding names of the experimental scales and the conventional clinical scale subscales that were among the significant findings of the studies reviewed.

(See Butcher & Tellegen, 1978, for additional recommendations concerning the methodological design of MMPI research in general.)

Review

The preceding methodological evaluation revealed several methodological shortcomings within the designs of the studies to be reviewed. As a result, the interpretation and generalization of the present findings are restricted. These studies can best be viewed as indicative of the potential general usefulness of the MMPI in many different areas of prison work rather than a final judgment of its value to corrections. The findings below will perhaps serve as starting points for better controlled studies, which should attempt to cross-validate and extend these results.

The present section reviews the studies in the following three main categories: The MMPI As a Primary Differentiator of Deviant Behavior, the MMPI As a Predictor of Deviant Behavior, and MMPI Differences Across Race and Sex. The first category is broken down into five sections: hostile-assaultive, first offenders versus recidivists, sexual deviancy, addictions, and classification of psychopathologic behavior. The second category is broken down into three sections: recidivism, prison escape, and institutional adjustment. Finally, the third category is broken down into two sections: racial differences and sex differences.

The MMPI As a Primary Differentiator of Deviant Behavior

Hostile-assaultive. The most extensive work in this field has centered around the development and validation of the Overcontrolled Hostility (*O-H*) scale by Megargee and his associates. Initially Megargee and Mendelsohn (1962; 2) attempted to differentiate between assaultive and nonassaultive criminals using 12 relevant experimental scales. No scale correctly isolated the assaultive group in the predicted manner, but the authors noticed one surprising trend: The protocols of the nonviolent group made them appear less controlled or more hostile than the aggressive groups. They suggested the following hypothesis:

> The extremely assaultive person is often a fairly mild-mannered, long-suffering individual who buries his resentment under rigid and brittle controls. Under certain circumstances he may lash out and release all his aggression in one, often disastrous, act. Afterwards he reverts to his usual overcontrolled defenses. Thus he may be more of a menace than the verbally aggressive, "chip-on-the-shoulder," type who releases his aggression in small doses. (p. 437)

Megargee, Cook, and Mendelsohn (1967; 2) explored this hypothesis in a complex study consisting of the generation of the 31-item *O-H* scale and two cross-validation attempts. They found that the scale seemed to identify the co-occurrence of two usually incompatible personality constructs, impulse control and hostile alienation. They felt that a high *O-H* scale score indicated "a conflict between strong aggressive impulses and strong inhibitions against the expression of aggression" (p. 528), which can manifest itself either in explosive violent acts or psychosis. They concluded that although the generation of an all-purpose assaultiveness scale from the MMPI seems unlikely, the *O-H* scale is capable of identifying a subgroup of assaultive criminals who are of this overcontrolled type. Megargee et al. (1967) did not advocate a specific cutting score. The best cutting score for their data identified 85.7% of the "extremely assaultive" type, producing 43.2% false positives.

Deiker (1974; 1) examined the *O-H* scale along with the conventional scales and 20 other experimental aggression scales on the MMPIs of four experimental groups: homicide, battery, threat, and a control group. Differences on 17 of the experimental scales were significant, but only 4 scales showed significance in the predicted direction: *O-H*, Ego Overcontrol (*Eo*), Direction of Hostility (*DH*), and Denial (*Dn*). Significant differences were also observed on the *F*, *K*, *Pd*, *Pt*, *Sc*, and *Ma* scales, with the control

group having the highest elevations on all except the *K* scale. Deiker concluded that although his results seemed to support Megargee (1966), a negative response bias hypothesis accounted for the results equally well. (The *O-H* scale has 21 items keyed false and 10 items keyed true.) Megargee and Cook (1975; 2) responded to this criticism by constructing two shortened and two lengthened *O-H* scales, all with equal numbers of scorable true and false items. They reanalyzed their initial set of protocols and found that one of the lengthened scales was actually a better discriminator than the original *O-H* scale, thereby refuting the negative-response bias hypothesis. Neither of these studies used cutting scores, instead they compared group means.

Davis and Sines (1971; 1) have discovered a profile configuration that seems to be associated with hostile-aggressive acting-out behavior. (These 4-3 studies are the sole exceptions to the prototypical research design included in this review, since their approach resembles that of Megargee et al., 1967. However, the importance and relevance of these findings demanded their inclusion in this review.) They concisely defined the 4-3 configurational prototype (profile peak on the *Pd* scale with the second highest elevation on the *Hy* scale) and examined the differences between these profiles and a control group's profiles in each of three settings: a state hospital, a prison, and a medical center. A behavioral pattern of hostile-aggressive outbursts in usually quiet men consistently emerged across the three settings. The authors pointed out that this behavioral pattern is similar to Megargee's *O-H* type and Gilberstadt and Duker's (cited in Davis and Sines, 1971) "4-type," and they concluded that the consistency of their findings both across their samples and with previous work constitutes strong evidence of the prototype's validity. They speculated that 4-3 profile types have a constitutional predisposition toward this behavior pattern: 4-3 types seem to be controlled by a cyclical internal mechanism that periodically causes acute emotional and behavioral disturbances and seems impervious to conventional treatment methods. Persons and Marks (1971; 1) successfully replicated these results, noting the significantly higher incidence of violent crimes by 4-3 types when compared to three of the most common MMPI code types in prison. Davis (1971; 1) also obtained similar results with a female inmate population. The nature of these studies precluded the determining of false-positive rates.

Two other studies were found that investigated aggression in prisoners. Blackburn (1968; 1) found that his "extremely assaultive" group scored significantly higher than his "moderately assaultive" group on the *L, K, R, Eo,* and *Dn* scales and significantly lower on the *F, Pd, Ma,* Extraversion (*Ex*), and General Hostility (*GH*) scales. Even though the *O-H* scale was not yet available to him for this study, Blackburn concluded that his findings supported Megargee's (Megargee, 1966) overcontrolled hostility hypothesis. Carroll and Fuller (1971; 1) compared nonviolent, violent, and sexual offenders and found that the nonviolent group appeared to be hostile and confused in thinking and displayed the most deviant profiles. These nonviolent subjects were found to be significantly higher than the other two groups on the *F, Sc,* and *Ma* scales. Their findings would also seem to support Megargee's hypothesis; it is unfortunate that the *O-H* scale was not included in their analysis.

The potential value of these important findings is self-evident. The identification of potentially violent inmates with the help of the MMPI would aid in treatment plans, administrative decisions, and parole considerations. As Megargee and Mendelsohn (1962) point out, the detection of assaultiveness of the overcontrolled variety can be a difficult task, since these types usually appear so passive and mild-mannered. The cyclical nature of the 4-3 prototype may also be initially deceptive. Even though not all violent criminals are identified by these two major indices, these indices seem to isolate those types who may not be seen as assaultive until a harmful outburst occurs. Although further cross-validation is needed, the already demonstrated discriminative powers of both the *O-H* scale and the 4-3 prototype are encouraging. However, investigation of false-positive rates for both in-

dices must be pursued. Comparisons of high *O-H* scale profiles and 4-3 profiles should be carried out to determine if they are measuring essentially the same thing.

First offenders versus recidivists. A few studies have attempted to differentiate first offenders from recidivists with MMPIs obtained after the recidivists had returned to prison. Dunham (1954; 1) found that the *D* and *Pd* scales were significantly higher for recidivists, whereas Stanton (1956; 2) found that the *Pd* and *Ma* scales were higher. Panton (1959a; 2) found no significant differences between first offenders and recidivists. Others who have not restricted themselves to the conventional MMPI scales have come up with somewhat more meaningful results. Panton (1962b; 1) found that recidivists were significantly higher on the *Pd, Ma,* and Prison Adjustment (*Ap*) scales and combined the *Pd* and *Ap* scales to form the new Habitual Criminal (*HC*) scale. However, the results of his cross-validation (which blocked on age and number of prior offenses) were not compelling, and he noted that the scale seemed to get less effective as the number of prior offenses decreased. Pierce's (1972a; 1) attempt at cross-validation of the *HC* scale succeeded in effectively differentiating first offenders from recidivists. (He did not select a cutoff score, however.) Adams (1976; 2) also reexamined the *HC* scale and found that it effectively differentiated recidivists from first offenders, as did its parent scales *Pd* and *Ap*. Flanagan and Lewis (1974; 2) found that offenders with juvenile records scored significantly higher than "absolute first offenders" on the *F, Pd, Pa, Sc,* and *Ma* scales and significantly lower on the Responsibility (*Re*) scale. Christensen and LeUnes (1974; 1) used the Prison Adjustment Scale (*PAS*) in addition to the conventional scales, but they failed to achieve any significant differentiations with respect to recidivism.

These results are of little direct predictive usefulness due to their post hoc nature. Observed differences within this paradigm cannot be equated with differences extant in future recidivists before their initial release from prison. The most consistently significant findings across these studies showed recidivists to have relatively higher evaluations on the *Pd* and *Ma* scales, but these are common peaks for prison populations that are frequently above a *T* score of 70. Therefore, these results seem to offer no dynamic insights that could be useful in treatment approaches. The only possible utility the *HC* scale might have is if it could predict recidivism before the fact. In short, little useful information seems to be contained in these studies. (More directly predictive studies are examined in the section on Recidivism.)

Sexual deviancy. Work in this area has centered around the identification of homosexuals. Panton (1960b; 2) found the conventional *Mf* scale to be ineffective for this purpose, so he generated the Homosexual (*Hsx*) experimental scale. This scale identified 81% of the homosexuals and 87% of the nonhomosexuals in his initial sample and 86% of the homosexuals and 81% of the nonhomosexuals in his cross-validation sample. Pierce (1972b; 1) found that the *Hsx* scale identified 94% of his "active homosexual" group, 100% of his "situational homosexual" group (heterosexual before coming to prison), and 100% of his comparison group. A second study by Pierce (1973; 1) that used the *Hsx* scale differentiated 81% of his active homosexual group and 81% of his situational homosexual group with a cutoff score of 9.8, and scores on the scale remained relatively stable on a retest 1 year later. Another study by Pierce (Note 4; 1) that used both the *Hsx* scale and the *Mf* scale found that mean scale score differences for both scales successfully differentiated his active homosexual group from his situational homosexual and nonhomosexual groups, whereas neither scale successfully differentiated the latter two groups from each other. However, Cubitt and Gendreau (1972; 1) failed to effectively differentiate their homosexual group with the *Hsx* scale, even though other scales did significantly differentiate (i.e., the *Hs, D, Hy, Mf, Pa,* and both of Manosevitz's abridged *Mf* scales [Manosevitz, 1970]). They noted that their homosexual group was significantly older and that *Hsx* scores demonstrated a significant positive correlation with age; this should have exaggerated the power of the

Hsx scale to detect homosexuals in their sample. Cubitt and Gendreau conclude that the validity of the *Hsx* scale is "limited" but that the *Mf* scale (and Manosevitz's abridged versions) effectively discriminates homosexuals from heterosexuals. Panton's (1978; 1) most recent study effectively discriminated prior-to-incarceration homosexuals from a baseline prison sample, finding that homosexuals scored significantly higher on the *Mf, Pa, Sc, Ma, Hsx,* and Frustration Tolerance Index (*FTI*; Beall & Panton, Note 5) scales. These results led Panton to conclude that homosexuals exhibit "an implied greater social alienation and weaker impulse control characterized by a more likely acting-out to stress and frustration than is indicated for the population sample as a whole" (p. 11). Panton also found significant differences among demographic variables and Sex Inventory (Thorne, 1966) scales. He offered the overall conclusion that the homosexual entering prison will probably pursue his homosexual inclinations but will not necessarily become sexually assaultive in these pursuits.

Two other studies were found that examined group differences on conventional MMPI scales. Oliver and Mosher (1968; 1) compared a group of heterosexuals with homosexual "inserters" and homosexual "insertees." There were no significant differences between the two homosexual groups (possibly due to a small *N*), but the insertees were significantly higher than the heterosexuals on the *Hs, Hy, Pd, Mf,* and *Pt* scales, whereas the inserters were significantly higher on the *F, Hs, D, Hy, Pd, Pa, Pt,* and *Sc* scales. McCreary (1975; 1) compared child molesters with previous offenses to first-offense child molesters and found that previous offenders scored significantly higher on the *Pd, Pd2, Hs, Hy,* and *Sc* scales.

Although the MMPI appears to be sensitive to differences between homosexuals and heterosexuals, no one specific indicator was consistently effective across all studies. Cubitt and Gendreau's (1972) study was conducted in a Canadian prison, and as a result cultural differences may partially explain the failure of the *Hsx* scale to differentiate in their sample. Manosevitz's abridged *Mf* scales need to be examined further with prison populations. A number of the previously mentioned studies (Oliver & Mosher, 1968; Panton, 1978; Pierce, 1972) discuss the several types of disciplinary problems centering around the homosexual inmate (including fighting, homosexual seduction and rape, and attempted suicides or escapes to avoid homosexual demands), which dramatize the need to identify types of homosexual inmates as soon as possible for effective and beneficial treatment approaches and administrative decisions. Due to a lack of truly accurate figures for the incidence of inmate homosexuality as well as the possible human expense resulting from false negatives in chance prediction, Meehl and Rosen's (1955) base-rate prescriptions seem to be inapplicable here. All of the above research was conducted after homosexual behavior had already been detected in the criterion groups, which seems to be a sensible approach considering the relative stability of homosexual behavior over time. However, more research of a longitudinal nature is needed, preferably combining MMPI protocols obtained on admission with careful follow-up procedures to maximize the predictive utility of the MMPI in coping with problems caused by inmate homosexuality. Distinctions between types of homosexuals such as those made by Oliver and Mosher could prove to be valuable in assisting future predictive studies to identify homosexual subtypes who may instigate some of the disciplinary problems previously mentioned. Other types of sexual deviancy (such as sexual offenders) have yet to be explored to any significant extent, but the MMPI may also prove to be useful in these areas.

Addictions. A few studies have examined drug addiction by limiting themselves to the conventional MMPI scales. Gendreau and Gendreau (1970; 1) failed to find any significant differences between their groups of heroin addicts and nonaddicts and concluded that the "addiction-prone" theory (which seeks to ascribe specific personality traits to addicts) is an inappropriate approach to the problem. Panton (1977; 2) compared drug dealers to drug abusers and found that dealers scored significantly higher on the *Ma,* Amorality (*Ma1*), Psychomotor Acceleration

(*Ma2*), and Ego Inflation (*Ma4*) scales and significantly lower on the *Hs, D, Hy, Pt,* and *Si* scales. He concluded that drug dealers are a more difficult management and treatment problem than drug abusers.

Cavior, Kurtzberg, and Lipton (1967; 1) compared heroin addicts to nonaddicts and generated the Heroin (*He*) experimental scale that when applied to two cross-validation samples, correctly classified 75% of the overall adult sample and 67% of the overall adolescent sample. As part of a larger study, Panton and Brisson (1971; 2) compared drug abusers with nonusers and found that the drug group scored significantly lower on the Aggravated Sex (*Asx*) scale, significantly higher on the *Hy, Pd, Mf, Sc, Ma, Pa, Ec,* and *HC* scales and significantly higher on the Need for Affection (*Hy2*), Lassitude-Malaise (*Hy3*), Familial Discord (*Pd1*), Authority Problems (*Pd2*), Social Alienation (*Pd4A*), Personal and Emotional Sensitivity (*Mf1*), Denial of Masculine Occupations (*Mf5*), Social Alienation (*Sc1A*), Lack of Ego Mastery-Cognitive (*Sc2A*), and Imperturbability (*Ma3*) subscales. These results tended to support the several findings they had made through their examination of their other sources of data. Panton and Brisson also generated the Drug Abuser (*DaS*) experimental scale, which identified 75.4% of the drug abusers and 81.4% of the nonusers in their initial sample and 75.8% of both groups in a cross-validation sample. Panton and Behre (1973; 1) compared drug addicts to abusers without addiction on the conventional MMPI scales, 11 experimental scales (not including *He* or *DaS*) and several demographic variables. Contrary to Panton and Brisson's findings, they found no significant differences on the MMPI (several demographic variables did differentiate, however). The authors concluded that

> the MMPI results . . . support the contention of Gendreau and Gendreau that there is no such diagnostic identity as the addition-prone personality, and that imprisoned narcotics addicts do not necessarily have unique personality traits, as measured by the MMPI which predispose them toward the special effects of heroin, or distinguish them from non-addicted imprisoned drug abusers. (p. 416)

Panton (1972; 1) also investigated the validity of three alcoholism scales in a prison population. In comparing a prison alcoholic group with a prison nonalcoholic group and a normal nonalcoholic group, he found that the Alcoholism (*Al*) scale successfully differentiated the prison alcoholics (65.8% hit rate) from the prison nonalcoholics (also 65.8% hit rate). The second Alcoholism (*Am*) scale only differentiated the normal group from the prison groups, and the third Alcoholism (*Ah*) scale failed to achieve any differentiation. Panton concluded that even the *Al* scale seems to be affected by a more general factor of sociopathy, however. As part of his larger study, Stanton (1956; see the section on First Offenders Versus Recidivists) compared alcoholics with narcotics addicts and nonaddicts and found that alcoholics scored significantly higher on the *Pd* scale than either addicts or nonaddicts. He found no differences between the addict and nonaddict groups.

These sparse results are difficult to compare, since the criteria for formation of the criterion groups vary widely across the studies. Furthermore, little practically useful predictive potential is apparent, and the addiction-prone personality theory has not received any support. More reliable and expedient means of identifying addicts are already at the disposal of correctional personnel, so the application of experimental MMPI scales seems unnecessary. The MMPI might be able to identify prognostically favorable signs that could aid in the treatment of addicts, but evidence of such indicators does not presently exist. In short, the MMPI has not been shown to offer any practically significant insights into the addiction problem that cannot be more reliably gained from other sources; a productive use of the MMPI in this area has yet to be demonstrated.

Classification of psychopathologic behavior. Some studies have investigated the extent to which the MMPI can differentiate between different types of psychopathologic behavior. Gynther (1962; 1) examined the performance of 12 crime category groups of court-referred patients on the *F* scale and concluded that larger *F* values indicate "sicker" subjects. Craddick (1962; 1) extracted a

group of psychopaths and a group of nonpsychopaths from a larger pool of prisoners on the basis of scores on a checklist of psychopathic characteristics that he filled out for each prisoner. He found that the psychopaths obtained significantly higher scores on the *Pd, Pt,* and *Ma* scales. Johnston and Cooke (1973; 1) separated one pool of prisoners into four pairs of mutually overlapping groups (aggressive vs. nonaggressive, maximum security vs. nonsecurity placement, escape precaution vs. no precaution, and alcoholic diagnosis vs. nonalcoholic diagnosis) on the basis of behavioral records and clinical judgments. They found no significant differences between any of these pairs on the *Ah, Ec, Hc,* or Recidivism (*Rc*) experimental scales.

Several studies have used and expanded on theoretical classifications like Lindesmith and Dunham's (cited in Randolph, Richardson, & Johnson, 1961) socialized versus individualized juvenile delinquent types. Randolph et al. (1961; 2) paraphrased Lindesmith and Dunham's descriptions of these types as follows: "The socialized criminal is one who commits crimes that are supported and prescribed by his culture, so that, by committing a crime, the criminal gains in status and recognition" (p. 293). On the other hand, "the individualized criminal . . . acts for reasons that are personal and private. He commits his crimes alone and, in theory, is a stranger to others who commit similar crimes" (p. 293). Furthermore, "the socialized criminal seems likely to be a rather normal person," whereas "the individualized criminal, at odds with his own primary group, seems likely to be an individual whose criminality is merely symptomatic of deeper psychological pressures" (p. 293). Randolph et al. compared groups of these two types on the conventional scales and found that although profile configurations were similar, the solitary delinquents scored significantly higher on all clinical scales except the *Ma* scale. Wilcock (1964; 1) added an aggressive socialized group to his samples approximating the above two groups (this new group was described as combining certain elements of the first two groups) but merely found that the individualized group scored significantly higher than the other two groups on the *Hs* and *Hy* scales. (A small *N* may partially account for this lack of significance.)

Shinohara and Jenkins (1967; 3) used Jenkins' classifications (the first two of which are highly similar to Lindesmith and Dunham's two basic types) of socialized versus unsocialized aggressive versus runaway delinquents. This latter type was characterized "by repeatedly running away from home overnight, by staying out late at night, by stealing *in the home* . . . and by stealing which is furtive rather than aggressive" (Shinohara & Jenkins, 1967, p. 157; italics theirs). They examined differences between these groups on the conventional scales and 8 experimental scales. They found that the unsocialized aggressive group scored significantly higher than the socialized group on the *F, Hs, D, Pd, Pa, Sc,* and Anxiety (*At*) scales, whereas the runaway group scored significantly higher than the socialized group on the *F, Hs, D, Hy, Pd, Mf, Pt, Sc, At,* and Asocial (*As*) scales. They also found that the runaway group scored significantly higher than the unsocialized aggressive group on the *Mf* scale but scored significantly lower on the *Pa* scale. They concluded that the socialized group showed much less psychopathology and that their delinquency was "typically adaptive goal-oriented motivation behavior," whereas the other two groups showed more psychopathology and their delinquency was "frustration behavior rather than adaptive behavior" (pp. 161–162). Tsubouchi and Jenkins (1969; 3) attempted to validate and extend these findings using the conventional scales and 10 experimental scales. They found that their unsocialized aggressive group scored significantly higher than the socialized group on the *F, Hs, D, Pd, Pt,* Emotional Maturity (*Em*), and *HC* scales, whereas the runaway group scored significantly higher than the socialized group on the *F, Pd, Sc, Pa,* and *Em* scales. They also found that the unsocialized aggressive group scored significantly higher than the runaway group on the *Hs* scale. They concluded that the findings of Shinohara and Jenkins were essentially supported. In addition, they generated an experimental scale (unnamed) that separated their "motiva-

tion" delinquents (the socialized group) from their "frustration" delinquents (the other two groups).

Gregory (1974; 1) took a unique approach to this problem by generating sequential linear-sums formulas to identify delinquents who had been classified as psychopathic, adjusted, or neurotic on the basis of clinical graduate student checklist ratings of commitment summaries. Of Gregory's overall sample, 63% was classifiable by these three formulas, which used T scores on standard scales (e.g., Psychopathic Index [PI] $= F + 2Pd + Ma - 2K - Pt - 2Si - 40$). Gregory suggested that this approach to classification may be more efficient in general than the three other approaches to profile typing that he inspected and discarded as inadequate. (The three types are clinically derived code types, two-point techniques, and the "D^2 method.") Although Gregory emphasized the need for replication, no other studies were found that addressed his findings.

The work based on theoretical classifications such as Lindesmith and Dunham's (cited in Randolph et al., 1961) seems to be the most promising approach in this area at this point. Although criteria for inclusion into the different criterion groups varied across the studies, three out of four studies found extensive significant differences that supported their theoretical hypotheses. The pursuit of these findings could be profitable in maximizing the effectiveness of differential treatment approaches to these various groupings. More stable prima facie indicators derived from the MMPI for these criminal types would be a desirable extension of these findings; the investigation of Tsubouchi and Jenkins' (1969) experimental scale and the generation of other experimental scales toward this end seems to be the most advisable approach. The possible presence of prognostic MMPI indicators for these groups should also be examined. Gregory's sequential linear-sums approach demands further investigation, both with respect to his specific findings and to the general utility of his linear-sums methodology. The work of Megargee (1977) and his associates appears to be a more extensive and comprehensive effort at this type of classificatory MMPI system, but perhaps Megargee's empirical approach could be productively interfaced with the theoretical and empirical approaches previously described.

The MMPI As a Predictor of Deviant Behavior

Recidivism. These studies are distinguished from those in the First Offenders Versus Recidivists section in that they compare the MMPIs of recidivists and nonrecidivists that were administered before each subject's initial release from prison. Instead of a post hoc analysis seeking possible treatment approaches, these studies concentrate on using the MMPI as a reliable predictor of future recidivists. The base-rate predictive guidelines of Meehl and Rosen (1955) are particularly relevant here. Frank (1971; discussed later) cites the following Task Force on Corrections (1967) findings: "The best current estimates indicate that, among adult offenders, 35 to 45 per cent of those released on parole are subsequently returned to prison" (p. 3). Therefore, any study in this area that attempts to claim any practical significant value must comfortably exceed the 65% overall hit rate that could be realized by merely predicting that no released inmate will return to prison. The rate of false positives (those identified by the MMPI as recidivists who are actually nonrecidivists) must also be weighed against any gains over this chance expectancy figure (which creates false negatives only).

Only one study was found that confined itself to the examination of conventional MMPI scales. Mack (1969; 3) found no significant differences between his group of recidivists and parole successes. Other studies have attempted to generate or employ experimental scales as predictive indicators of recidivism. As part of a larger study, Gough et al. (1965; 1) examined the conventional scales and the Anxiety (A), Repression (R), and Ego Strength (ES) scales, and found that only the Ma scale was significantly higher for recidivists. Panton (1962c; 2) generated the 26-item Parole Violator (PaV) scale that identified 80.5% of the violators

and 80.5% of the nonviolators in his initial sample (an overall hit rate of 80.5%, with 12.7% false positives) and 78.6% of a cross-validation sample consisting solely of violators. He also found that nonviolators scored significantly lower than violators on the *Hs*, *D*, *Hy*, *Pd*, *Pa*, *Pt*, *Sc*, and *Ma* scales and significantly higher on the *Mf* scale.

A major work in this area is the dissertation of Black (1967; 2). He examined 15 experimental scales as well as the conventional scales and found that no single scale achieved a significant differentiation between his groups of recidivists and nonrecidivists. However, he took two of the scales (*Si* and *HC*) that had the highest correlations with the criterion, selected items that the two groups had responded to differentially five or more times, eliminated overlapping items, and called the results the Recidivism–Rehabilitation scale (*Rmn*; inspection reveals that this scale does not share any items with Panton's *PaV* scale). This 22-item scale initially identified 88% of the recidivists and 84% of the nonrecidivists in his sample. Attempting to improve on these results, Black then constructed the *Rmn* index, which yields a score of 1 point for each of the following scale elevation criteria: *Si* scale *T* score less than 54, *HC* scale *T* score greater than 58, *Rmn* scale *T* score greater than 50. Scores of 0 and 1 on this index were found to predict rehabilitation, and a score of 3 was found to predict recidivism, with a score of 2 being indecisive. However, Black found that the arithmetical difference between the *A* and *R* experimental scales successfully identified 80% of the two-score recidivists with differences of 8 or less and 75% of the two-score nonrecidivists with differences greater than 8. This combined system achieved an overall predictive accuracy of 90%, with only a 7.8% rate of false positives. Black organized all of his findings into the Recidivism–Rehabilitation Inventory, which presents the following scales in a format modeled after the conventional MMPI profile sheet: *L*, *F*, *K*, *Si*, *HC*, *Rmn*, *A*, *R*, *PaV*, *Pd*, *Ma*, *Ec*, and *ES*.

A dissertation by Frank (1971; 2) sought to cross-validate these findings. He found that the *Rmn* scale alone was the best predictor for his sample, identifying 75.5% of the parole successes and 68% of the recidivists (a 73.1% overall hit rate, with 15.9% false positives). The *Rmn* index identified 80.0% of the successes and 54.4% of the recidivists (an overall hit rate of 71.0% with 13% false positives), and Frank found that the major cause of misclassifications seemed to be the *A* scale minus *R* scale difference for *Rmn* index scorers of two. Frank points out that Black's subjects were tested just prior to release, whereas his subjects were tested on their initial admission to prison, and he admits that the failure of his testing procedure to include the effects of incarceration may have compromised the predictive powers of the *Rmn* scale and index. He found that no single item of the *Rmn* scale was a more efficient discriminator than the scale as a whole, and this led him to tentatively conclude that "the Recidivism–Rehabilitation scale taps a recidivist 'syndrome,' a personality dimension with overt MMPI response tendencies" (pp. 37, 39).

Two studies reviewed elsewhere in this article make passing reference to the prediction of recidivism. Wattron (1963; see the section on Institutional Adjustment) found that his 72-item Prison Maladjustment Scale (*PM*) identified 68% of his recidivists and 69% of his successful parolee group (an overall hit rate of 68.7%, with 20.1% false positives. Inspection reveals only a four-item overlap between the *PM* and *Rmn* scales: Items 56, 118, 216, and 469 all scored true). Davis and Sines (1971; see the Hostile-Assaultive section) noted that the 12 men in their sample with 4–3 profiles who were subsequently paroled all violated parole within 1 year of release.

The only study that showed an appreciable gain over chance prediction in percentage of overall hit rate with a corresponding smaller percentage of false positives was Black's (1967) work on the *Rmn* scale and index. The potential value of these findings cannot be assessed at this time, but they demand attempts at cross-validation, since even the shrinkage in accuracy that is expected would still allow a significant improvement over chance prediction. Black's work also includes an extensive review of the

literature on recidivism, and he effectively integrates his findings with the theories and research findings of others in the field. The implications for both the successful identification of future recidivists and the formulation of more effective treatment approaches for this particular group are far-reaching indeed. Frank's (1971) attempted cross-validation of Black's findings was disappointing, but it was compromised by the difference from Black's procedure in the timing of MMPI administration; it seems reasonable to assume that MMPIs administered just prior to release would probably be better discriminators of recidivism than MMPIs given on initial admission to prison. (In fact, Table 2 reveals that Black's study was the only study reviewed that administered all MMPIs just prior to release.) Also, success by Frank would not have greatly extended the generalizability of the *Rmn* scale and index anyway because he conducted his replication in the same state as Black's original work (Oklahoma). Further cross-validations that replicate Black's methodology more precisely must be conducted in other parts of the country before the value of Black's findings can be adequately assessed.

Panton's *PaV* scale showed some gain in percentage of accurate prediction while producing a corresponding smaller percentage of false positives; future studies may also profit by considering the *PaV* scale along with Black's *Rmn* scale and index. Attempts at cross-validation must incorporate the vital requirement of adequate follow-up to minimize contamination of their nonrecidivist sample with "recidivists-to-be." Frank cites evidence that indicates that over 80% of recidivists return to prison within 2 years of release, so it would seem that a thorough 2-year follow-up would be a minimum requirement for such studies. In summary, it would seem that the predictive powers of the MMPI for the identification and treatment of potential recidivism with the use of the *Rmn* scale and index appear to be promising but need further study.

Prison escape. Once again, Meehl and Rosen's (1955) criteria constitute the focal point of this evaluation. Shupe and Bramwell's (1963; see the section on Methods of Profile Interpretation) 5% escape figure will be used to estimate the effectiveness of predicting escape risk, since the studies discussed in this section did not supply escape base rates for their particular prison populations. Therefore, a successful escape predictor should exceed a 95% overall hit rate with a minuscule incidence of false positives. This is a tall order indeed.

All work in this area has centered around the 42-item *Ec* scale that was generated by Beall and Panton (1956; 1) from a sample tested after escape attempts had already occurred. This scale identified 76.7% of the escapees and 73% of the nonescapees (an overall hit rate of 73.2%, with 25.6% false positives) in the original sample and 77.2% of the escapees and 78.3% of the nonescapees (an overall hit rate of 78.2%, with 20.6% false positives) in the cross-validation sample. Pierce (1971b; 1) employed alternative cutoff scores for two pairs of groups, the first pair with a criterion group having a record of escape prior to the MMPI administration and the second pair with a criterion group having attempted escape subsequent to MMPI administration. Using the cutoffs that achieved the greatest dichotomy, the *Ec* scale identified 80% of the prior escapees and 82% of the nonescapees (an overall hit rate of 81.9%, with 17.1% false positives) in the first pair, whereas in the second pair the *Ec* scale identified 50% of the future escapees and 76% of the nonescapees (an overall hit rate of 74.7%, with 22.8% false positives).

As part of a larger study, Stump and Gilbert (1972; 1) obtained mean scores on the *Ec* scale for a group attempting escape prior to the MMPI and a group attempting escape subsequent to the MMPI and compared these means to the mean *Ec* scale score obtained by the general prison population. The group attempting escape prior to the MMPI had an *Ec* scale mean that was significantly higher than the general population mean, but the other criterion group was not significantly different. Adams and West (1976; 2) found no significant differences between group *Ec* scale means for either of their two comparisons, one in-

volving inmates with one escape attempt subsequent to the MMPI versus a no-escape group and the other involving inmates with two or more escape attempts subsequent to the MMPI versus a no-escape group. Panton (Note 6; 1; Note 7; 2) ran two comparisons with admissions MMPIs, one involving escapees versus nonescapees and the other involving groups with three or more escapes, two escapes, one escape, and no escapes. In the first comparison, the *Ec* scale identified 74% of the escapees and 70% of the nonescapees (an overall hit rate of 70.2%, with 28.5% false positives), whereas in the second comparison the scale identified 90.4% of the three-or-more-escapes group, 86.7% of the two-escapes group, 80.3% of the one-escape group, and 71.4% of the no-escape group (an overall hit rate of 72%, with 27.2% false positives).

Even if the corresponding base escape rate was much higher than the 5% rate assumed here, the *Ec* scale would not appreciably improve prediction over chance expectancy. Unless a high rate of false positives is acceptable for security grade assignment, no benefit is apparent from the scale. Meehl and Rosen (1955) suggest that if a subpopulation can be isolated that has an appreciably higher base rate of the criterion than the overall population in question, indices that lack utility in this overall population may prove to be useful. The classifications of Megargee (1977) and his associates might help in this respect. Perhaps examining the number of escape attempts would be more appropriate than examining the number of actual escapes; this would enlarge the base rate and possibly make the *Ec* scale a more useful tool. However, the *Ec* scale does not seem to have any practical utility at the present time.

Institutional adjustment. Several studies have examined groups of disciplinary problem inmates by focusing primarily on the conventional MMPI scales. Driscoll (1952; 1) created four groups of prisoners ranging on a continuum from "most maladjusted" to "most adjusted." He found that the most adjusted group was significantly higher than any of the other groups on the *D, Mf,* and *Pa* scales and that the most maladjusted group presented the most normal MMPI profiles. This led Driscoll to speculate that the prison environment fosters modes of adaptive behavior that are viewed as maladaptive behaviors outside of prison. Erikson and Roberts (1966; 3) initially compared a maladjusted juvenile delinquent group to an adjusted group, and they found that the maladjusted group scored significantly higher on the *Pd* scale. They generated a 19-item scale that differentiated the two groups, but two replication attempts failed to support any of their initial findings. Lefkowitz (1966; 1) compared adjustment "failures" with adjustment "successes" and found that the failure group scored significantly higher on the *Ma* scale. Twomey and Hendry (1969; 2) compared discipline problem inmates with a comparison group and found that the discipline problem group scored significantly higher on the *L, F, Hs, D, Hy, Mf, Pa, Pt, Sc,* and *Si* scales. Snortum, Hannum, and Mills (1970; 1) rated a group of women offenders on a continuum representing frequency of rule violations and discovered that frequency of rule violations positively correlated with elevations on the *Pd* and *Ma* scales. Two of these studies (Driscoll, 1952, and Snortum et al., 1970) obtained their MMPIs before the disciplinary problems occurred, whereas the rest were obtained after the research groups had been formed.

None of the predictive studies discussed later considered the base-rate probability of their criterion variable, which was most commonly some indicator of disciplinary actions taken. Although these base-rate statistics are probably not routinely gathered, Meehl and Rosen (1955) suggest that a simple analysis of available records would produce a suitable base rate for comparative purposes. It is recommended that future studies rectify this observed methodological flaw.

Predictive studies have investigated the ability of experimental scales to identify potential disciplinary problems. Panton (1958; 2) generated the 36-item Prison Adjustment Scale (*Ap*) and examined its discriminative powers on the profiles of two adjusted inmate groups, two nonadjusted groups and

one severely nonadjusted group. The scale correctly identified 82% of each adjusted group, 87% and 85% of the two nonadjusted groups, and 93% of the severely nonadjusted group. Pierce (1972c; 1) compared inmates with two or more infractions to a comparison group and found that Panton's cutoff scores identified 88% of the maladjusted inmates but produced a false-positive rate of 50%. Edwards (1963; 1) compared groups of first-offender juvenile "successes," first-offender juvenile "failures," and prison inmate "failures" on the conventional scales, the Harris-Lingoes subscales (Harris & Lingoes, Note 8), 14 experimental scales (including *Ap*), and mean number of high-point elevations. He found that the prison inmate failure group scored significantly higher than the juvenile groups on the *Sc* scale and the *Sc2A* subscale but found no other differences. (This may be partially due to a small *N*.)

Wattron (1963; 1) compared parole and maladjusted inmates and generated the Prison Maladjustment Scale (*PM*). This scale successfully identified 82% of the maladjusted group and 84% of the parolees in his cross-validation sample. Stump and Gilbert's (1972; see the section on Prison Escape) previously cited study compared the *Ap* scale scores of a group of inmates repeatedly disciplined by solitary confinement with the *Ap* scale scores of a group that had spent no nights in solitary confinement, but this produced no significant differentiation. Panton (1973; 2) compared a group of management problem inmates with a large baseline inmate sample and examined the conventional scales, his Prison Classification Inventory (PCI; Panton, Note 9), modeled after the conventional MMPI profile sheet, consisted of the following conventional and experimental MMPI scales: *L, F, K, Ap, Ec, HC, PaV, Hsx, A, R, Pd*, Defect of Inhibition Control (*Dc&i*), Sensorimotor Dissociation (*SD*), and *Asx*, the Harris-Lingoes subscales and five additional experimental scales. He found that the management problem inmates scored significantly higher than the baseline sample on the *F, Pd, Pd2, Mf, Ma, Ma1, Ap, Ec, HC*, and *Re* scales, and these inmates scored significantly lower on the *Hs, Pt, Si, A*, and *R* scales. Panton noted that the management problem inmates exhibited uniformly poor prognostic MMPI signs (high scores on the *Pd, Ma*, and *HC* scales and low scores on the *Hs, D, Pt*, and *A* scales), and he concluded that their susceptibility to rehabilitative efforts "appears limited."

Sutker and Moan (1973; 1) examined differences on the conventional scales and the *Ap* and *PM* experimental scales between two successive pairings of groups of "bad actors" (severe discipline-problem inmates) and "no disciplines." Their first group of bad actors scored significantly higher than the no disciplines on the *F, Ma*, and *PM* scales, whereas the second group of bad actors scored significantly higher on the *F, Hy*, and *Ma* scales. They noted that both no-disciplne groups actually scored higher on the *Ap* scale than their bad-actors counterparts, although the differences were not significant. Two of these studies (Panton, 1973; Sutker & Moan, 1973) obtained MMPI profiles after the criterion and comparison groups were formed, whereas the rest of the studies obtained their MMPIs before the discipline problems occurred.

One study was found that examined self-mutilating prisoners. Panton (1962a; 1) compared a group of self-mutilators with a group of model prisoners and a group of infraction-nonmutilators on the conventional scales and eight experimental scales. The infraction-nonmutilators were matched with the self-mutilator group "as to number and type of infraction and degree of exposure to custodial stress and pressure" (p. 63). The infraction-nonmutilator group scored significantly higher than the model prisoners group on the *Ap* scale only. The self-mutilator group scored significantly higher than the other two groups on the *F, Pa, Pt, Sc*, Anxiety Index (*AI*), Critical Item (*CI*), Inner Madadjustment (*In*), Judged Manifest Hostility (*Jh*), and Psychotic Tendency (*Pq*) scales and significantly lower on the Conversion Reaction (*CR*) scale. The self-mutilator group also scored significantly higher than the model prisoners group on the *Pd* and *Ap* scales. Panton concluded that "the self-mutilators were more inclined toward

compulsive outbursts of hostility, appeared more anxious, expressed a greater inner turmoil, and appeared more inclined toward bizarreness in their overt resistance to stringent attempts to control their aggressiveness" (p. 66). Panton then generated the Self-Mutilator scale (*SM*) by adding the *T* scores of the *F*, *Pa*, and *Sc* scales together and determining a cutoff score. This scale identified 83.8% of the self-mutilators, 75.6% of the model prisoners, and 81.2% of the infraction-nonmutilators in his original sample.

In this latter study, Panton (1962a) notes that with respect to the self-mutilators, "none of the group claimed they were actually attempting to destroy themselves nor did the psychiatric examinations reveal any evidence in support of suicidal intentions" (p. 63). No other studies were found that even made reference to suicidal prisoners. This lack of suicide research in prisons is difficult to explain, especially since attempted suicides constitute a dangerous problem for both administrative and treatment personnel. Generalization of suicide research results from other populations (e.g., mental hospitals) is an unwarranted procedure, and in any case such generalization is no substitute for findings that are based on prison populations. This important issue demands exploratory research on the MMPI profiles of suicidal prisoners, since the ability to effectively identify suicidal risks in an economical manner would obviously be of great value to prison personnel. Although the overall incidence of suicidal attempts is probably small, Meehl and Rosen (1955) point out that their base-rate considerations sometimes cease to be relevant, especially in certain "life-and-death" situations; surely a moderate rate of false positives in suicidal risk prediction is an acceptable price to pay for the trade-off in preserved human lives.

Those studies that focused on the conventional MMPI scales have only come up with inconsistent and inconclusive results. This may be partially due to different criteria for the formation of criterion groups, different timing of MMPI administrations, and the different populations that were examined. More well-controlled MMPI research may yet offer valuable dynamic insights into the reasons behind maladjusted behavior in some prison inmates. Panton's *Ap* scale has not yet proven useful in other parts of the country, although he has consistently obtained positive results with this scale in North Carolina; again, differences across studies with respect to timing of MMPI administration, criteria for the formation of criterion groups, and different types of prison populations may partially account for this lack of consistency. Wattron's *PM* scale needs further research, since the only attempted cross-validation obtained significant results in just one of two comparisons. Other potentially valuable indices such as the Harris-Lingoes subscales, Panton's PCI, and Panton's *SM* scale also need further investigation before any conclusions about their worth can be drawn. This entire area merits further study, since effective predictors of potential prison adjustment problems would be invaluable to both administrative and treatment personnel. Groups differentiated along such dimensions could be placed within the prison so as to minimize potential disturbances, and treatment plans individually geared to such differentiated groups could bring an increase in positive rehabilitation results. Future research should attempt to cross-validate existing MMPI indicators and should consider the generation of new experimental scales that may prove to be more effective than existing ones. The lack of research on the MMPIs of suicidal risk prisoners constitutes a glaring omission in this research area that should be rectified immediately.

MMPI Differences Across Race and Sex

Racial differences. Several studies have examined the differences in conventional scale elevations produced by groups of black and white inmates. Stanton's (1956) previously cited study (see the section on First Offenders Versus Recidivists) initially compared groups of black and white inmates and found no significant differences. Caldwell (1959; 2) pursued a similar design and found that black inmates obtained significantly higher scores on the *Hs*, *D*, *Mf*, *Pa*,

and *Ma* scales and significantly lower scores on the *Pd* scale. Panton's (1959a) previously cited study (see the section on First Offenders Versus Recidivists) found that black inmates scored significantly higher than white inmates on the *F, Pa, Sc,* and *Ma* scales and significantly lower on the *Hy* scale. As part of a larger study, Costello, Fine, and Blau (1973; 1) found no significant differences between their samples of black and white inmates. Elion and Megargee (1975; 3) investigated the validity of the *Pd* scale among black males using groups of prisoners and groups of college students. They found that black inmates scored significantly higher on the *Pd* scale than both a group of "culturally deprived black male university students" and a group of white inmates. They concluded that elevations on the *Pd* scale validly express levels of social deviance among young black males but that the present scale norms appear to show racial bias.

A few studies have examined black–white differences with respect to experimental scales. Panton (1959b; 2) compared black and white inmates on the Harris-Lingoes subscales and found that black inmates scored significantly higher on the Ideas of External Influence (*Pa1*), *Sc1A,* and *Ma4* subscales and significantly lower on the *Pd2* subscale. Haven (Note 10; 1) investigated racial differences on Megargee's *O-H* scale and found that black inmates scored significantly higher than white inmates, suggesting either that black inmates as a group experience more feelings of social alienation or that they have been shaped by societal pressure to more actively inhibit aggression. (Haven favored the latter alternative.) As part of a larger study, Fisher (1969; 1) investigated racial differences on the Repression-Sensitization (*R-S*) scale and found that white inmates scored significantly higher than their black counterparts, suggesting that the black inmates showed evidence of more repression.

More recent studies have directly confronted the issue of racial and cultural bias in the MMPI with their comparisons of prisoners from different racial and cultural groups. McCreary and Padilla (1977) compared 40 black, 36 Mexican American, and 267 white male misdemeanor offenders who had been convicted and were awaiting sentencing. They compared individual scale elevations as well as scores on Goldberg's (1965) linear classification system and ran unmatched comparisons as well as comparisons in which subjects were matched on educational level and occupation. They hypothesized that differences due to socioeconomic factors would appear only in the unmatched comparisons, whereas differences due to cultural factors would appear in both comparisons. Analyses were conducted using only valid profiles (less than 30 items left blank, *F* scale less than a raw score of 23, and the *F-K* index less than 11) and using all profiles, but there were virtually no differences in the results of these two analyses, so valid profile results were used to report most findings. Mexican Americans had significantly less education than whites, but no other significant differences emerged with respect to educational level or occupation. In the unmatched comparisons, blacks were significantly higher on the *Al* scale and significantly lower on the *Hy* and *Mf* scales than whites. Mexican Americans were significantly higher on the *L, Hs,* and *O-H* scales than whites. (They were also higher on the *K* scale when all profiles were used.) In the matched comparisons, blacks scored significantly higher on the *K* and *Ma* scales and significantly lower on the *Hy* scale than whites. (Only the *Ma* scale difference remained significant when all profiles were used.) The Mexican Americans scored higher on the *L, K,* and *O-H* scales than whites. The only significant difference on the Goldberg indices showed that on the psychiatric–sociopathic index, the Mexican Americans scored in the psychiatric range, whereas the whites scored in the sociopathic range (unmatched condition). McCreary and Padilla concluded that both cultural and socioeconomic factors seemed to contribute to the observed MMPI differences between these three groups.

Rosenblatt and Pritchard (1978) compared 104 black and 191 white male inmates on the MMPI with respect to differences on full-scale WAIS IQ scores. They rejected number of years in school as an appropriate indicator of educational achievement because

most of their sample of Mississippi inmates had been educated in segregated schools that were not considered to be equivalent in educational quality. They divided their sample into four subgroups with respect to race and IQ (using the overall sample's mean IQ of 93 to divide the groups). They found no racial differences between the high-IQ groups but found that the low-IQ blacks ($N = 81$) scored significantly higher on the *Hs*, *Sc*, and *Ma* scales and significantly lower on the *Hy* scale than the low-IQ whites ($N = 70$). Successive applications of more stringent validity rules did not affect these findings. Rosenblatt and Pritchard concluded that racial differences on the MMPI seem to be limited to low-IQ subjects.

The findings on the conventional scales are not definitive at this time. The most consistent finding appears to be a higher elevation for blacks than whites on the *Ma* scale (five studies, including Panton's Harris-Lingoes subscales findings), with a similar but weaker trend on the *Sc* and *Pa* scales (three studies each, all of which included Panton's Harris-Lingoes subscale findings) and the *Hs* scale (two studies). Blacks also seem to occasionally score lower than whites on the *Hy* scale (three studies) and the *Mf* scale (two studies). Only Elion and Megargee (1975) found significantly higher elevations for black inmates on the *Pd* scale, whereas significant differences in two other studies (including Panton's Harris-Lingoes subscale findings) showed white inmates scoring significantly higher on the *Pd* scale. None of the findings for the experimental scales have been replicated. The most crucial findings here appear to be the results of Rosenblatt and Pritchard (1978), which tend to support the position on racial bias in the MMPI (discussed in the Sources of Variance and Their Effects on Test Results section). Apparent racial bias in the MMPI may actually be due to educational factors, since more intelligent blacks do not seem to display the differences in MMPI performance that less intelligent groups of blacks display. More research is definitely needed to resolve the issue of racial and cultural bias in the MMPI, and such efforts should investigate the differences between black and white MMPI inmate protocols with respect to education and/or IQ.

Sex differences. Research with the MMPI in prison populations has been conducted almost completely with male inmates. Only two of the studies reviewed previously (Davis, 1971; Snortum et al., 1970) used female inmate populations. Although there are far fewer female inmates than male inmates in the United States today, the female inmate population "presents a significant minority whose needs in terms of proper classification, treatment and training are as great as the needs of their more numerous male counterparts" (Panton, 1974). Two studies were found that compared samples of male and female prison inmates. Panton (1974; 2) examined differences on conventional scales and the Harris-Lingoes *Pa* subscales and found that female inmates scored significantly higher than male inmates on the Poignancy (*Pa2*) and *Si* scales and scored significantly lower on the *Hs* and *D* scales. Panton concludes that male inmates "appear to be more anti-social with neurotic overlays," whereas the female inmates "appear more asocial than anti-social with overlays of greater emotional sensitivity" (p. 332). Joesting, Jones, and Joesting (1975; 1) examined differences on the conventional scales and seven experimental scales and found that female inmates scored significantly lower than male inmates on the *F*, *Hs*, *D*, *Hy*, *Pd*, *Mf*, *Pa*, *Sc*, *Ma*, *Si*, *Ec*, *A*, *R*, *Pd1*, *Dc&i*, and *SD* scales, and scored significantly higher on the *K* and *Ap* scales. They concluded that the males appeared more emotionally disturbed than the females. Although certainly lacking in consistency, both of these studies suggest significant MMPI protocol differences between male and female prison inmates that preclude the generalizability of the research findings in this article to female inmate populations. Attempted replication of these male inmate findings with female inmate populations is the only possible solution to this problem.

Panton (1976; 2) compared a group of male prisoners admitted in 1966 to a group of male prisoners admitted in 1971 to investigate any population changes over time on the *Mf* scale and the Pepper and Strong

(Note 11) *Mf* subscales. He found that the 1971 group scored significantly higher on the *Mf*, *Mf1*, and *Mf5* scales, and concluded that the 1971 sample showed "a greater personal and emotional sensitivity and a more frequent rejection of masculine occupations and avocations" (p. 606) than their 1966 counterparts. This finding needs replication but suggests that considerable caution should be employed in applying any of the findings concerning the *Mf* scale that were discussed in this review. Although the essential meanings of these findings are probably not affected, any *T*-score levels employed for interpretive purposes may be anachronistic and therefore misleading.

Conclusions and Recommendations

The studies covered in this review can most accurately be described as a beginning effort in the attempt to maximize the usefulness of the MMPI in corrections. As mentioned in the introduction, these findings are best viewed as indicative of the potential of the MMPI in prison work, not as a final judgment of its worth. Since the overall representativeness of the experimental samples employed was often restricted, and several methodological shortcomings were apparent, the generalizability of the present findings is limited. The issue of racial and cultural bias in the MMPI with respect to prison populations in particular is a vital question that demands extensive research in it's own right, since at present one cannot determine with an acceptable degree of certainty just what a minority group member's MMPI means and does not mean. No legitimate inferences can be generalized from the present findings to women inmates, and the effects of age and IQ will probably modify interpretations somewhat when the nature of these effects is more precisely defined. A strong note of caution is necessary with respect to the appropriate use of this test: The MMPI serves as a source of probabilistic clinical statements that are hypotheses for further exploration, and employment of the MMPI as the sole basis for any kind of decision effecting the life of a subject constitutes a serious abuse of the test. This is because the "noise" inherent in any test of this type always creates the danger of false positive conclusions, and as a result the MMPI should be used in conjunction with other sources of data to form the basis of important decisions affecting a subject's welfare if it is to be employed in an enlightened and ethical manner. In any event, these initial research findings do appear to be promising. Further research is necessary for a conclusive assessment of the MMPI's potential for becoming an important and valuable aid in correctional practice that would benefit the inmate as well as the correctional personnel responsible for effective program functioning.

The MMPI shows potential in several areas. The findings concerning the *O-H* experimental scale and the 4-3 prototype possess a strong preliminary research base that needs cross-validation; however, evidence is sufficiently strong at this point to warrant the provisional application of these indicators in correctional practice as warning signs of potentially dangerous violent outbursts. Several other findings appear promising but need further study. For example, the identification of homosexuality with the MMPI could be profitably employed in administrative and treatment decisions after some further exploration and refinement of the present findings. The classifications of psychopathologic behavior described previously could also be of great use to administrative and treatment personnel if expanded by further study. The *Rmn* experimental scale and index demands instant examination, for it could prove to be a highly effective predictor of recidivism that could substantially improve hit rates over chance prediction. Finally, the pursuit of the findings of Megargee (1977) and his associates may prove to be more fruitful than can be presently imagined. Inconclusive findings exist on indicators of institutional adjustment, but the MMPI could yet prove to be worthwhile in this area. The identification of suicide risks in prison has been inexplicably ignored to date and demands immediate exploration. Areas investigated in which the MMPI appears to be of little present worth include the post hoc comparison of first offenders versus recidivists and

the identification of addictions and escape risks. In short, further research on the MMPI's use in several of these areas could produce results that would prove to be invaluable to correctional personnel.

In 1970, Haven (Note 1) stated that "an overview of all the research reviewed gives the rather discouraging picture of a hodgepodge of one-shot investigations. There were few follow-ups and little cross-evaluation of previous findings" (p. 39). Unfortunately, this still seems to be largely true 7 years later. With the two outstanding exceptions of Panton and Megargee, there is no evidence of any concerted effort at comprehensive longitudinal study with thorough follow-up procedures for the MMPI's use in prison work. Such efforts are sorely needed and can be instigated by the routine testing of prison admissions with the full 566-item MMPI to build up a data bank. However, inspection of Table 2 reveals that approximately two-thirds of the studies reviewed seem to have drawn their MMPIs from just such a data bank; the lack of well-controlled longitudinal research becomes increasingly difficult to justify in the light of this fact. A few selected areas such as recidivism may be more profitably investigated with MMPIs administered long after initial admission, yet the comparison of such protocols with the corresponding admissions protocols across the criterion and comparison groups could in itself provide dependable indices of specific inmate behaviors. In summary, more longitudinal research with thorough follow-up procedures is needed if effective investigation of the MMPI's potential in corrections is to be realized.

Past methodological shortcomings must be eliminated if future results are to prove fruitful. Future research must observe either standard random sampling procedures or standard matching procedures if their findings are to be representative; the presence of a large MMPI data bank should facilitate the meeting of these requirements. Investigators must also instigate better controls or post hoc checks for variables such as age, race, and IQ to maximize the representativeness and generalizability of their results. In fact, these and several other design aspects (such as *F* scale elevations, fake-good and fake-bad indicators, and the several remaining subject characteristics listed in Table 2) demand further research in their own right to produce methodological refinements that should increase the MMPI's sensitivity to significant differences. One of the most important improvements that needs to be made over previous predictive research is the incorporation of Meehl and Rosen's (1955) essential guidelines prescribing the consideration of base-rate probabilities. The different approaches to MMPI data interpretation discussed in the Methodology section should also be kept in mind, since the simultaneous investigation of more than one of these approaches may produce unexpectedly valuable findings. Once again, longitudinal research with thorough follow-up procedures is a must. Future researchers who incorporate these important guidelines into their experimental designs will maximize both their ability to detect actual significant differences and the generalizability of their findings.

The central purpose of this review has been to stimulate needed cross-validation of existing findings as well as to encourage further exploratory research with the MMPI in prison work. In addition, it is hoped that the interfacing of these findings with the ongoing exploration of the new classification system of Megargee (1977) and his associates will enhance the precision and productivity of the MMPI's employment in corrections. The full extent of the MMPI's ultimate usefulness in this respect cannot yet be fully appreciated, but its potential should not be underestimated at this point. The MMPI may someday prove to be an indispensable factor in the creation of more effective rehabilitative approaches to correctional practice.

Reference Notes

1. Haven, H. The MMPI with incarcerated adult and delinquent offenders. *FCI Technical and Treatment Notes*. Tallahassee, Fla.: Federal Correctional Institute, 1970.
2. Grayson, H. M. *A psychological admissions testing program and manual*. Unpublished manuscript, Veteran's Administration Center, Neuropsychiatric Hospital, Los Angeles, 1951.

3. Megargee, E. I. Personal communication, January 8, 1978.
4. Pierce, D. M. Differences between active and situational homosexuality on two MMPI scales. *ISR Research Bulletin* (1974, *8*, 1–15). Anamosa, Ia.: Iowa State Reformatory.
5. Beall, H. S., & Panton, J. H. *Development of the prison adjustment scale (PAS) for the MMPI.* Unpublished manuscript, 1957. (Available from J. H. Panton, North Carolina Department of Correction, 840 West Morgan Street, Raleigh, N.C. 27603.)
6. Panton, J. H. *Research note on the re-investigation of the MMPI escape scale.* Unpublished manuscript, 1974. (Available from the North Carolina Department of Correction, 840 West Morgan Street, Raleigh, N.C. 27603.)
7. Panton, J. H. *Validation of the MMPI escape scale* (Ec). Unpublished manuscript, 1977. (Available from the North Carolina Department of Correction, 840 West Morgan Street, Raleigh, N.C. 27603.)
8. Harris, R. E., & Lingoes, J. C. *Subscales for the MMPI: An aid to profile interpretation.* Unpublished manuscript, Department of Psychiatry, University of California School of Medicine, Los Angeles, 1955.
9. Panton, J. H. *Development of a prison classification inventory (PCI) for the MMPI* (Rev. ed.). Unpublished manuscript, Department of Correction, Raleigh, N.C., 1970.
10. Haven, H. Racial differences on the MMPI *O-H* scale. *FCI Research Reports* (1969, *1*(5), 1–9). Tallahassee, Fla.: Federal Correctional Institute.
11. Pepper, L. J., & Strong, P. N. *Judgmental subscales for the* Mf *scale of the MMPI.* Unpublished manuscript, Department of Health, Honolulu, Hawaii, 1958.

References

Adams, T. C. Some MMPI differences between first and multiple admissions with a state prison population. *Journal of Clinical Psychology,* 1976, *32,* 555–558.

Adams, T. C., & West, J. E. Another look at the use of the MMPI as an index to "escapism." *Journal of Clinical Psychology,* 1976, *32,* 580–582.

Beall, H. S., & Panton, J. H. Use of the MMPI as an index to escapism. *Journal of Clinical Psychology,* 1956, *4,* 392–394.

Bennett, L. Test-taking "insight" of prison inmates and subsequent parole adjustment. *Correctional Psychologist,* 1970, *4,* 27–34.

Black, W. G. The description and prediction of recidivism and rehabilitation among youthful offenders by the use of the MMPI[2] (Doctoral dissertation, University of Oklahoma, 1967). *Dissertation Abstracts,* 1967, *28,* 1691B. (University Microfilms No. 67-11, 996)

Blackburn, R. Personality in relation to extreme aggression in psychiatric offenders. *British Journal of Psychiatry,* 1968, *114,* 821–828.

Blumberg, S. MMPI *F* scale as an indicator of severity of psychopathology. *Journal of Clinical Psychology,* 1967, *23,* 96–99.

Buechley, R., & Ball, H. A new test of "validity" for the MMPI. *Journal of Consulting Psychology,* 1952, *16,* 299–301.

Butcher, J. N., & Tellegen, A. Common methodological problems in MMPI research. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 620–628.

Caldwell, M. G. Personality trends in the youthful male offender. *Journal of Criminal Law, Criminology and Police Science,* 1959, *49,* 405–416.

Carroll, J. L., & Fuller, G. An MMPI comparison of three groups of criminals. *Journal of Clinical Psychology,* 1971, *27,* 240–242.

Cavior, N., Kurtzberg, R. L., & Lipton, D. S. The development and validation of a heroin addiction scale with the MMPI. *The International Journal of the Addictions,* 1967, *2,* 129–137.

Christensen, L., & LeUnes, A. Discriminating criminal types and recidivism by means of the MMPI. *Journal of Clinical Psychology,* 1974, *30,* 192–193.

Comrey, A. L. Factor analysis of the *F* scale. *Educational and Psychological Measurement,* 1957–1958, *18,* 621–623.

Costello, R. M., Fine, H. J., & Blau, B. I. Racial comparisons on the MMPI. *Journal of Clinical Psychology,* 1973, *29,* 63–65.

Costello, R. M., Tiffany, D. W., & Grier, R. H. Methodological issues and racial (black–white) comparisons on the MMPI. *Journal of Consulting and Clinical Psychology,* 1972, *38,* 161–168.

Cowan, M., Watkins, B., & Davis, W. Level of education, diagnosis and race-related differences in MMPI performance. *Journal of Clinical Psychology,* 1975, *31,* 442–444.

Craddick, R. A. Selection of psychopathic from non-psychopathic prisoners within a Canadian prison. *Psychological Reports,* 1962, *10,* 495–499.

Cronbach, L. J., & Gleser, G. L. *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press, 1965.

Cubitt, G. H., & Gendreau, P. Assessing the diagnostic utility of the MMPI and 16 PF indexes of homosexuality in a prison sample. *Journal of Consulting and Clinical Psychology,* 1972, *39,* 342.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook: Vol. 1. Clinical interpretation.* Minneapolis: University of Minnesota Press, 1972.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook: Vol. 2. Research applications.* Minneapolis: University of Minnesota Press, 1975.

Davis, K. R. The actuarial development of a female 4'3 MMPI profile (Doctoral dissertation, Saint Louis University, 1971). *Dissertation Abstracts International,* 1971, *32,* 1207B. (University Microfilms No. 71-28, 378)

Davis, W., & Jones, M. Negro versus caucasian psychological test performance revisited. *Journal of Consulting and Clinical Psychology,* 1974, *42,* 675–679.

Davis, K. R., & Sines, J. An antisocial behavior pattern associated with a specific MMPI profile. *Journal of Consulting and Clinical Psychology,* 1971, *36,* 229–234.

Deiker, T. A cross-validation of MMPI scales of aggression on male criminal criterion groups. *Journal of Consulting and Clinical Psychology,* 1974, *42,* 196–202.

Driscoll, P. Factors related to the institutional adjustment of prison inmates. *Journal of Abnormal and Social Psychology,* 1952, *47,* 593–596.

Dunham, R. E. Factors related to recidivism in adults. *Journal of Social Psychology,* 1954, *39,* 77–91.

Edwards, J. A. Rehabilitation potential in prison inmates as measured by the MMPI. *Journal of Criminal Law, Criminology and Police Science,* 1963, *54,* 181–185.

Elion, V., & Megargee, E. I. Validity of the MMPI *Pd* scale among black inmates. *Journal of Consulting and Clinical Psychology,* 1975, *43,* 166–172.

Erikson, R. V., & Roberts, A. H. An MMPI comparison of two groups of institutionalized delinquents. *Journal of Projective Techniques and Personality Assessment,* 1966, *30,* 163–166.

Fisher, G. The repression–sensitization scale: Effects of several variables and two methods of obtaining scores. *Journal of General Psychology,* 1969, *80,* 183–187.

Flanagan, J., & Lewis, G. First prison admissions with juvenile histories and absolute first offenders: Frequencies and MMPI profiles. *Journal of Clinical Psychology,* 1974, *30,* 358–360.

Foster, V. H., & Goddard, H. H. The Ohio literacy test. *Pedagogical Seminary,* 1924, *31,* 340–351.

Frank, C. H. The prediction of recidivism among young adult offenders by the recidivism–rehabilitation scale and index (Doctoral dissertation, University of Oklahoma, 1970). *Dissertation Abstracts International,* 1971, *32,* 557B. (University Microfilms No. 71-17, 042)

Gauron, E. R., Stevenson, R., & Englehart, R. MMPI *F* scores and psychiatric diagnosis. *Journal of Consulting Psychology,* 1962, *26,* 488.

Gendreau, P., & Gendreau, L. P. The "addiction-prone" personality: A study of Canadian heroin addicts. *Canadian Journal of Behavioral Science,* 1970, *2*(1), 18–25.

Gendreau, P., Irvine, M., & Knight, S. Evaluating response set styles on the MMPI with prisoners: Faking good adjustment and maladjustment. *Canadian Journal of Behavioral Science,* 1973, *5,* 183–194.

Glenn, R. *A study of personality patterns of male defective delinquents as indicated by the MMPI.* Unpublished master's thesis, Pennsylvania State University, 1949.

Goldberg, L. R. Diagnosticians versus diagnostic signs: The diagnosis of psychosis versus neurosis from the MMPI. *Psychological Monographs,* 1965, *79* (9, Whole No. 602).

Gough, H. G., Wenk, E. A., & Rozynko, V. V. Parole outcome as predicted from the CPI, the MMPI, and a base expectancy table. *Journal of Abnormal Psychology,* 1965, *70,* 432–441.

Gregory, R. Replicated actuarial correlates for three MMPI code types in juvenile delinquency. *Journal of Clinical Psychology,* 1974, *30,* 390–394.

Gynther, M. D. The clinical utility of "invalid" MMPI *F* scores. *Journal of Consulting Psychology,* 1961, *25,* 540–542.

Gynther, M. D. Crime and psychopathology. *Journal of Abnormal and Social Psychology,* 1962, *64,* 378–380.

Gynther, M. D. White norms and black MMPIs: A prescription for discrimination? *Psychological Bulletin,* 1972, *78,* 386–402.

Gynther, M. D., Altman, H., & Warbin, R. W. Interpretation of uninterpretable MMPI profiles. *Journal of Consulting and Clinical Psychology,* 1973, *40,* 78–83.

Gynther, M. D., Lachar, D., & Dahlstrom, W. G. Are special norms for minorities needed? Development of an MMPI *F* scale for blacks. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 1403–1408.

Gynther, M. D., & Petzel, T. P. Differential endorsement of MMPI *F* scale items by psychotics and behavior disorders. *Journal of Clinical Psychology,* 1967, *23,* 185–188.

Gynther, M. D., & Shimkunas, A. M. Age, intelligence, and MMPI *F* scores. *Journal of Consulting Psychology,* 1965, *29,* 383–388. (a)

Gynther, M. D., & Shimkunas, A. M. More data on $F > 16$ MMPI profiles. *Journal of Clinical Psychology,* 1965, *21,* 275–277. (b)

Hathaway, S. R., & McKinley, J. C. *Minnesota Multiphasic Personality Inventory manual.* New York: The Psychological Corporation, 1967.

Jastak, J., Bijou, S., & Jastak, S. *Wide Range Achievement. Test: Manual.* New York: Psychological Corporation, 1965.

Joesting, J., Jones, N., & Joesting, R. Male and female prison inmates' differences on MMPI scales and Revised Beta IQ. *Psychological Reports,* 1975, *37,* 471–474.

Johnston, N., & Cooke, G. Relationship of MMPI alcoholism, prison escape, hostility control and recidivism scales to clinical judgments. *Journal of Clinical Psychology,* 1973, *29,* 32–34.

Kent, G. H. *Series of emergency scales.* New York: Psychological Corporation, 1946.

Kincannon, J. C. Prediction of the standard MMPI scale scores from 71 items: The Mini-Mult. *Journal of Consulting and Clinical Psychology,* 1968, *32,* 319–325.

Lawton, M. P., & Kleban, M. H. Prisoners' faking on the MMPI. *Journal of Clinical Psychology*, 1965, *21*, 269–271.

Lefkowitz, M. M. MMPI scores of juvenile delinquents adjusting to institutionalization. *Psychological Reports*, 1966, *19*, 911–914.

Mack, J. L. The MMPI and recidivism. *Journal of Abnormal Psychology*, 1969, *74*, 612–614.

McCreary, C. P. Personality differences among child molesters. *Journal of Personality Assessment*, 1975, *39*, 591–593.

McCreary, C. P., & Padilla, E. MMPI differences among black, Mexican-American and white male offenders. *Journal of Clinical Psychology*, 1977, *33*, 171–177.

McKegney, F. P. An item analysis of the MMPI *F* scale in juvenile delinquents. *Journal of Clinical Psychology*, 1965, *21*, 201–205.

Manosevitz, M. Item analysis of the MMPI *Mf* scale using homosexual and heterosexual males. *Journal of Consulting and Clinical Psychology*, 1970, *35*, 395–399.

Meehl, P. E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 1955, *52*, 194–216.

Megargee, E. I. Undercontrolled and overcontrolled personality types in extreme antisocial aggression. *Psychological Monographs*, 1966, *80*(3, Whole No. 611).

Megargee, E. I. The need for a new classification system. *Criminal Justice and Behavior*, 1977, *4*, 107–114. (a)

Megargee, E. I. Directions for further research. *Criminal Justice and Behavior*, 1977, *4*, 211–216. (b)

Megargee, E. I., & Bohn, M. J. Empirically determined characteristics of the ten types. *Criminal Justice and Behavior*, 1977, *4*, 149–210.

Megargee, E. I., & Cook, P. E. Negative response bias and the MMPI *O-H* scale: A response to Deiker. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 725–729.

Megargee, E. I., Cook, P. E., & Mendelsohn, G. A. Development and validation of an MMPI scale of assaultiveness in overcontrolled individuals. *Journal of Abnormal Psychology*, 1967, *72*, 519–528.

Megargee, E. I., & Dorhout, B. Revision and refinement of the classificatory rules. *Criminal Justice and Behavior*, 1977, *4*, 125–148.

Megargee, E. I., & Mendelsohn, G. A. A cross-validation of twelve MMPI indices of hostility and control. *Journal of Abnormal and Social Psychology*, 1962, *65*, 431–438.

Meyer, J., & Megargee, E. I. Initial development of the system. *Criminal Justice and Behavior*, 1977, *4*, 115–124.

Morrice, J. K. W. The MMPI in recidivist prisoners. *Journal of Mental Science*, 1957, *103*, 632–635.

Oliver, W. A., & Mosher, D. L. Psychopathology and guilt in heterosexual and subgroups of homosexual reformatory inmates. *Journal of Abnormal Psychology*, 1968, *73*, 323–329.

Panton, J. H. Predicting prison adjustment with the MMPI. *Journal of Clinical Psychology*, 1958, *14*, 308–312.

Panton, J. H. Inmate personality differences related to recidivism, age and race as measured by the MMPI. *Journal of Correctional Psychology*, 1959, *4*, 28–35. (a)

Panton, J. H. The response of prisoners to MMPI subscales. *Journal of Social Therapy: Corrective Psychiatry*, 1959, *5*, 233–237. (b)

Panton, J. H. MMPI code configurations as related to measures of intelligence among a state prison population. *Journal of Social Psychology*, 1960, *51*, 403–407. (a)

Panton, J. H. A new MMPI scale for the identification of homosexuality. *Journal of Clinical Psychology*, 1960, *16*, 17–21. (b)

Panton, J. H. The identification of predispositional factors in self-mutilation in a state prison population. *Journal of Clinical Psychology*, 1962, *18*, 63–67. (a)

Panton, J. H. The identification of habitual criminalism with the MMPI. *Journal of Clinical Psychology*, 1962, *18*, 133–136. (b)

Panton, J. H. Use of the MMPI as an index to successful parole. *Journal of Criminal Law, Criminology and Police Science*, 1962, *53*, 484–488. (c)

Panton, J. H. A validity study of three MMPI scales measuring alcoholism. *Correctional Psychologist*, 1972, *5*, 160–166.

Panton, J. H. Personality characteristics of management problem prison inmates. *Journal of Community Psychology*, 1973, *1*, 185–191.

Panton, J. M. Personality differences between male and female prison inmates measures by the MMPI. *Criminal Justice and Behavior*, 1974, *1*, 332–339.

Panton, J. H. Significant increases in MMPI *Mf* scores within a state prison population. *Journal of Clinical Psychology*, 1976, *32*, 604–606.

Panton, J. H. Personality characteristics of aged inmates within a state prison population. *Offender Rehabilitation*, 1976–1977, *1*(2), 203–208.

Panton, J. H. Personality characteristics of drug pushers incarcerated within a state prison population. *Quarterly Journal of Corrections*, 1977, *1*, 11–13.

Panton, J. H. Characteristics associated with male homosexuality within a state correctional population. *Quarterly Journal of Corrections*, 1978, *2*, 26–31.

Panton, J. H., & Behre, C. Characteristics associated with drug addiction within a state prison population. *Journal of Community Psychology*, 1973, *4*, 411–416.

Panton, J. H., & Brisson, R. C. Characteristics associated with drug abuse within a state prison population. *Corrective and Social Psychiatry and Journal of Behavior Technology Methods and Therapy*, 1971, *17*(4), 3–33.

Penk, W. E., & Robinowitz, R. MMPI differences of black and white drug abusers. JSAS *Catalog of Selected Documents in Psychology*, 1974, *4*, 50. (Ms. No. 630)

Persons, R., & Marks, P. The violent 4-3 MMPI personality type. *Journal of Consulting and Clinical Psychology*, 1971, *36*, 189–196.

Pierce, D. M. A cross-validation of the MMPI habitual criminalism scale. *Correctional Psychologist*, 1971, *4*, 183–187. (a)

Pierce, D. M. The escapism scale of the MMPI as a predictive index. *Correctional Psychologist*, 1971, *4*, 230–232. (b)

Pierce, D. M. MMPI correlates of adaptation to prison. *Correctional Psychologist*, 1972, *5*, 43–47. (a)

Pierce, D. M. MMPI *Hsx* scale differences between active and situational homosexuals. *Journal of Forensic Psychology*, 1972, *4*, 31–38. (b)

Pierce, D. M. Prison adjustment: Cross-validation of an MMPI scale. *Correctional Psychologist*, 1972, *5*, 22–24. (c)

Pierce, D. M. Test and nontest correlates of active and situational homosexuality. *Psychology*, 1973, *10*(4), 23–26.

Randolph, M. H., Richardson, H., & Johnson, R. C. A comparison of social and solitary male delinquents. *Journal of Consulting Psychology*, 1961, *25*, 293–295.

Rice, D. G. Rorschach responses and aggressive characteristics of MMPI $F > 16$ scorers. *Journal of Projective Techniques and Personality Assessment*, 1968, *32*, 253–261.

Rosenblatt, A. I., & Pritchard, D. A. Moderators of racial differences on the MMPI. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 1572–1573.

Shinohara, M., & Jenkins, R. L. MMPI studies of three types of delinquents. *Journal of Clinical Psychology*, 1967, *23*, 153–163.

Shupe, D. R., & Bramwell, P. F. Prediction of escape from MMPI data. *Journal of Clinical Psychology*, 1963, *19*, 223–226.

Snortum, J. R., Hannum, T. E., & Mills, D. H. The relationship of self-concept and parent-image to rule violations in a women's prison. *Journal of Clinical Psychology*, 1970, *26*, 284–287.

Stanton, J. M. Group personality profile related to aspects of antisocial behavior. *Journal of Criminal Law and Criminology*, 1956, *47*, 340–349.

Stump, E. S., & Gilbert, W. W. Experimental MMPI scales and other predictors of institutional adjustment. *Correctional Psychologist*, 1972, *5*(3), 141–154.

Sutker, P. B., & Moan, C. E. Prediction of socially maladaptive behavior within a state prison system. *Journal of Community Psychology*, 1973, *1*, 74–78.

Task Force on Corrections. *Task force report: Corrections* (The President's Commission on Law Enforcement and Administration of Justice). Washington, D.C.: U.S. Government Printing Office, 1967.

Thorne, F. C. The sex inventory. *Journal of Clinical Psychology*, 1966, *22*, 375–378.

Tsubouchi, K., & Jenkins, R. L. Three types of delinquents: Their performance on the MMPI and PCR. *Journal of Clinical Psychology*, 1969, *25*, 353–358.

Twomey, J. F., & Hendry, C. H. MMPI characteristics of difficult-to-manage federal penitentiary offenders. *Psychological Reports*, 1969, *24*, 546.

Wattron, J. B. A prison maladjustment scale for the MMPI. *Journal of Clinical Psychology*, 1963, *19*, 109–110.

Wilcock, K. D. Neurotic differences between individualized and socialized criminals. *Journal of Consulting Psychology*, 1964, *28*, 141–145.

Received February 22, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 964–973

Validity Conditions in Repeated Measures Designs

Huynh Huynh and Garrett K. Mandeville
College of Education
University of South Carolina

This article has two objectives. The first is to present necessary and sufficient conditions for the validity of traditional within-subject F tests in repeated measures designs. It is shown that the Mauchly sphericity criterion (W) and possibly the Box test for the equality of covariance matrices are appropriate to judge the validity of these conditions. Valid applications of both tests are conducted on sets of orthogonal normalized variables that are associated with each cluster of within-subject mean square ratios. The second objective of the article is to present empirical results on the appropriateness of using the W criterion when the variates are not normally distributed. For light-tailed distributions, the W criterion was shown to be moderately conservative, whereas for heavy-tailed distributions, empirical Type I error rates exceeded nominal alpha. Since most social science applications typically involve light-tailed rather than heavy-tailed distributions, the W criterion should provide useful results in most cases.

Traditional univariate analyses of variance for repeated measures (or mixed model) designs are used extensively in educational and psychological research (Kirk, 1968; Winer, 1971). In most situations observations are made on each subject in the sample under each combination of conditions (the within-subject factors). For example, a developmental psychology study may require measurements at several time intervals, with alternate forms used at each point in time. Thus, Time and Form would constitute two within-subject factors. If no factors that differentiate subjects are included, the design implies the testing of three within-subject hypotheses dealing with the effects of Time, Form, and the Time × Form interaction. Corresponding to each of these effects is a (nonunique) set of orthogonal normalized (orthonormal) variables.

On the other hand, if the subjects are stratified on one or more independent (between-subjects) factors, the interaction of these factors and the repeated measures factors provides more within-subject hypotheses that are subject to testing. In general, however, all within-subject hypotheses may be grouped into clusters that are tested using a common error term. In the example, if subjects were categorized according to age level, the mean squares associated with Time and the Age × Time interaction would be tested against the Time × Subject-Within-Age error term. The second cluster of hypotheses deals with Form and the Age × Form interaction, and the last group focuses on the interactions Time × Form and Time × Form × Age. As in the case of no independent factors, each of these clusters of effects and the error terms against which they are tested are associated with a set of orthonormal variables. These variables remain unchanged regardless of whether the subjects are subdivided into independent categories or not. As is discussed later, within normal distributions only the orthonormal variables will play a role in the validity of the traditional univariate F tests for repeated measures designs similar to the layout in the developmental psychology example.

A confusion exists regarding the validity conditions for the within-subject F tests in

Requests for reprints should be sent to Huynh Huynh, College of Education, University of South Carolina, Columbia, South Carolina 29208.

repeated measures designs. Originally these tests were derived from models implying equal variances and equal covariances (the compound symmetry condition) for the repeated measures (Scheffé, 1959). Several textbooks and articles written in the 1960s or earlier (Collier, Baker, Mandeville, & Hayes, 1967; Kirk, 1968) tended to treat compound symmetry as the required (e.g., necessary and sufficient) condition for the validity of the within-subject F tests. Huynh and Feldt (1970) and Rouanet and Lepine (1970), however, have shown that compound symmetry is only a sufficient condition. Although counterexamples are illustrated in Huynh and Feldt (1970), authors of more recent textbooks (Ferguson, 1976; Keppel, 1973), articles (Davidson, 1972; Poor, 1973), and computer manuals (Nie & Hull, Note 1) still mistake compound symmetry as a sine qua non assumption for the F tests.

Confusion also persists in statements regarding the validity conditions for designs with between-subjects (independent) factors. Huynh and Feldt (1970, Theorems 2 and 4) prove that these conditions involve only the orthonormal variables and give a counterexample, showing that it is unnecessary to assume equality of covariance matrices for the repeated measures across the independent factors. The latter condition is quoted in several design textbooks (Keppel, 1973; Winer, 1971, p. 523). To resolve this confusion, the next section of this article discusses the conditions under which the traditional within-subject F tests in repeated measures designs are valid.

Description of the Validity Conditions

The validity conditions under normality for the within-subject F tests have been investigated by Huynh and Feldt (1970) for randomized block (one-factor repeated measures) designs and for simple split-plot designs (two-factor designs with repeated measures on one factor), by Rouanet and Lepine (1970) for randomized block designs, by Mendoza, Toothaker, and Cain, (1976) for three-factor designs with repeated measures on two factors, and by Huynh (1978) for complex designs involving both independent and repeated factors. In the most general terms, the mean square ratios in each cluster of hypotheses follow exact F distributions if and only if (a) the covariance matrices for the associated set of orthonormal variables are identical across all levels of the independent factors, and (b) the common covariance matrix has a sphericity pattern (i.e., equal variances and zero covariances). If there are no independent factors, then the first condition does not apply. It may be noted that both conditions are based on the orthonormal variables and not on the original repeated measures. Therefore, they are more general than the requirements of equality of the covariance matrices for the original repeated measures and of compound symmetry for the common matrix. The remainder of this section displays the necessary and sufficient conditions for four typical situations.

Case 1: One-Factor Designs With Repeated Measures on the Factor

Assume that all subjects are measured under the b levels of the repeated factor B, and let the vector $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_b)'$ be the observation vector. Let $\mathbf{M}$ be any $(b-1) \times b$ matrix of $(b-1)$ orthonormal row vectors. Then $\mathbf{Y} = \mathbf{MX}$ transforms the b original variables to $(b-1)$ orthonormal variables. For example, if $b = 4$, the matrix $\mathbf{M}$ may be taken as follows:

$$\mathbf{M} = \begin{bmatrix} 1/(2)^{\frac{1}{2}} & -1/(2)^{\frac{1}{2}} & 0 & 0 \\ 1/(6)^{\frac{1}{2}} & 1/(6)^{\frac{1}{2}} & -2/(6)^{\frac{1}{2}} & 0 \\ 1/(12)^{\frac{1}{2}} & 1/(12)^{\frac{1}{2}} & 1/(12)^{\frac{1}{2}} & -3/(12)^{\frac{1}{2}} \end{bmatrix}.$$

Let $\Sigma(\mathbf{X})$ be the covariance matrix of $\mathbf{X}$. Then the covariance matrix of $\mathbf{Y}$ is $\Sigma(\mathbf{Y}) = \mathbf{M}\Sigma(\mathbf{X})\mathbf{M}'$. Within normal distributions, the mean square ratio for the B effects follows an exact F distribution if and only if $\Sigma(\mathbf{Y}) = \lambda\mathbf{I}_{b-1}$, $\mathbf{I}_{b-1}$ being the identity matrix of order $(b-1)$. In other words, the $(b-1)$ orthonormal variables $\mathbf{Y}$ are independent and equally variable, that is, the sphericity pattern holds.

Case 2: Two-Factor Designs With Repeated Measures on One Factor

Suppose that the subjects in Case 1 are categorized into a levels of an independent

Table 1
Matrices Defining the Orthogonal Normalized Variables for the B, C, *and* BC *Mean Square Ratios*

	Repeated measures								
Matrix	*B*1*C*1	*B*1*C*2	*B*1*C*3	*B*2*C*1	*B*2*C*2	*B*2*C*3	*B*3*C*1	*B*3*C*2	*B*3*C*3
	B within-subject effects								
$\mathbf{M}_B =$	$1/(6)^{1/2}$	$1/(6)^{1/2}$	$1/(6)^{1/2}$	$-1/(6)^{1/2}$	$-1/(6)^{1/2}$	$-1/(6)^{1/2}$	0	0	0
	$1/(18)^{1/2}$	$1/(18)^{1/2}$	$1/(18)^{1/2}$	$1/(18)^{1/2}$	$1/(18)^{1/2}$	$1/(18)^{1/2}$	$-2/(18)^{1/2}$	$-2/(18)^{1/2}$	$-2/(18)^{1/2}$
	C within-subject effects								
$\mathbf{M}_C =$	$1/(6)^{1/2}$	$-1/(6)^{1/2}$	0	$1/(6)^{1/2}$	$-1/(6)^{1/2}$	0	$1/(6)^{1/2}$	$-1/(6)^{1/2}$	0
	$1/(18)^{1/2}$	$1/(18)^{1/2}$	$-2/(18)^{1/2}$	$1/(18)^{1/2}$	$1/(18)^{1/2}$	$-2/(18)^{1/2}$	$1/(18)^{1/2}$	$1/(18)^{1/2}$	$-2/(18)^{1/2}$
	BC within-subject interaction effects								
$\mathbf{M}_{BC} =$	1/2	−1/2	0	−1/2	1/2	0	0	0	0
	$1/(12)^{1/2}$	$-1/(12)^{1/2}$	0	$1/(12)^{1/2}$	$-1/(12)^{1/2}$	0	$-2/(12)^{1/2}$	$2/(12)^{1/2}$	0
	$1/(12)^{1/2}$	$1/(12)^{1/2}$	$-2/(12)^{1/2}$	$-1/(12)^{1/2}$	$-1/(12)^{1/2}$	$2/(12)^{1/2}$	0	0	0
	1/6	1/6	−2/6	1/6	1/6	−2/6	−2/6	−2/6	4/6

(between-subjects) factor A. The within-subject hypotheses regarding the B and AB effects share the same error term. Thus, they belong to one cluster, and the corresponding orthonormal variables are defined by the **Y** vector of Case 1. Let $\Sigma_i(\mathbf{Y})$, $i = 1, \ldots, a$ be the covariance matrices of **Y** for the a levels of Factor A. The B and AB mean square ratios follow exact F distributions if and only if (a) the $\Sigma_i(\mathbf{Y})$ matrices are identical, and (b) the common covariance matrix has the sphericity pattern.

Case 3: Two-Factor Designs With Repeated Measures on Both Factors

Consider now a $B \times C$ two-factor design in which both B and C are repeated factors. This corresponds to the developmental psychology study described previously in which all subjects are pooled together in one group. There are three within-subject mean square ratios—one for the B effects, one for the C effects, and one for the BC interactions. Each corresponds to a set of orthonormal variables, namely $\mathbf{Y}_B$ for B, $\mathbf{Y}_C$ for C, and $\mathbf{Y}_{BC}$ for BC. (See Table 1 for an illustration.) Mendoza et al. (1976) show that each mean square ratio has an exact F distribution if and only if the corresponding orthonormal variables are independent and equally variable. Hence, the F test for B is valid if and only if $\Sigma(\mathbf{Y}_B) = \lambda_B \mathbf{I}_{b-1}$. The F test for C requires that $\Sigma(\mathbf{Y}_C) = \lambda_C \mathbf{I}_{c-1}$, and the F test for BC demands that $\Sigma(\mathbf{Y}_{BC}) = \lambda_{BC} I_{(b-1)(c-1)}$. In these conditions, the constants λ_B, λ_C, and λ_{BC} need not be equal.

Case 4: Three-Factor Designs With Repeated Measures on Two Factors

Finally let $A \times B \times C$ be a three-factor design in which A is the independent (between-subjects) factor and both B and C are repeated (within-subject) factors. The within-subject mean square ratios are then grouped into the following three clusters: the B and AB effects, the C and AC effects, and the BC and ABC effects. Each cluster corresponds to a set of orthonormal variables defined as in Case 3 (e.g., as if there were no independent factor). Let $\Sigma_i(\mathbf{Y}_B)$, for example, be the covariance matrix of the $\mathbf{Y}_B$ orthonormal variables at the ith level of factor A. Then the two mean square ratios for B and AB follow exact F distributions if and only if (a) the $\Sigma_i(\mathbf{Y}_B)$ matrices are identical, and (b) the common covariance matrix has a sphericity pattern. Similar necessary and sufficient conditions hold for the other two clusters of mean square ratios.

Testing for Sphericity

As indicated in each case of the previous section, the validity of the within-subject F tests requires sphericity for some covariance matrix. The matrix may be either the covariance matrix of a suitably chosen set of orthonormal variables or the common covariance matrix, if the design has some between-subjects factor(s). Thus, a test for sphericity is required if preliminary testing is to be considered.

Let the p-component vector $\mathbf{Y}$ be normally distributed with unknown mean vector and covariance matrix $\Sigma(\mathbf{Y})$. The assumptions of independence and variance homogeneity for the p components of $\mathbf{Y}$ are equivalent to the condition that $\Sigma(\mathbf{Y}) = \lambda \mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix of order p. Our interest is in testing the above condition as a null hypothesis (H_0) against the alternative hypothesis H_1: $\Sigma(\mathbf{Y}) \neq \lambda \mathbf{I}_p$. Let $\Sigma(\mathbf{Y})$ be estimated by the sample covariance matrix $\mathbf{S}$ based on a random sample of n vectors. Then the likelihood ratio test for H_0 against H_1 is of the form $W = |\mathbf{S}|/|\text{trace } \mathbf{S}/p|^p$ (Mauchly, 1940). The exact sampling distribution of W has been provided by Consul (1967, 1969), Pillai and Nagarsenker (1971), Mathai and Rathie (1970), and Nagarsenker and Pillai (1972, 1973; these also provide tables of critical values for W).

In the context of repeated measures designs, the matrix $\mathbf{S}$ is the sample covariance matrix of each set of orthonormal variables if there are no independent factors (Cases 1 and 3). For designs with independent factor(s), $\mathbf{S}$ is taken as the pooled covariance matrix if the assumption of equal covariance matrices for the orthonormal variables is tenable (Cases 2 and 4).

Testing for Equality of Covariance Matrices

As illustrated in Cases 2 and 4, suitably chosen sets of orthonormal variables must share the same covariance matrix across the independent factors for the corresponding within-subject mean square ratios to follow exact F distributions. Within normal distributions, equality of covariance matrices may be tested via the Box modified likelihood ratio criterion M (Morrison, 1976; Timm, 1975; Winer, 1971). Let p be the number of orthonormal variables and k be the number of levels of the independent factor. Let $\mathbf{S}_i$ be the traditional unbiased estimate of Σ_i, associated with the ith level of the independent factor and based on n_i degrees of freedom. Then the hypothesis $\Sigma_1 = \ldots = \Sigma_k$ may be tested via the Box criterion:

$$M = \Sigma n_i ln|\mathbf{S}| - \sum_{i=1}^{k} n_i ln|\mathbf{S}_i|,$$

where

$$\mathbf{S} = \sum_{i=1}^{k} n_i \mathbf{S}_i / \Sigma n_i$$

is the pooled estimate of the common covariance matrix. If

$$\rho = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\left(\Sigma \frac{1}{n_i} - \frac{1}{\Sigma n_i}\right),$$

then ρM approximately follows a chi-square distribution with $(k-1)p(p+1)/2$ degrees of freedom if the condition of equality of the covariance matrices holds. Tables of critical values of M at the .05 level may be found in Korin (1969) or Pearson and Hartley (1972) for the case of equal sample size.

To test the validity conditions in the presence of independent factors, a two-step procedure is suggested. First, the Box test may be carried out for suitably chosen sets of orthonormal variables. If equality of covariance matrices is tenable, then the Mauchly test for sphericity may be carried out on the pooled estimate $\mathbf{S}$ of the common covariance matrix. If α is the (joint) probability of the Type I error of the two-step procedure, then each step may be performed at the $\alpha/2$ level of significance. The two following sections present numerical illustrations for Cases 3 and 4.

Numerical Examples

Example 1

Table 2 presents the basic data for 22 subjects in a $B \times C$ design with repeated measures on both B and C. Table 1 displays the

Table 2
Basic Data for Numerical Examples 1 and 2

	Repeated measures								
Subject	B1C1	B1C2	B1C3	B2C1	B2C2	B2C3	B3C1	B3C2	B3C3
1	53	20	12	14	42	30	10	5	63
2	23	55	77	10	2	30	56	30	50
3	20	30	50	43	12	30	53	21	20
4	3	20	77	45	53	32	65	30	20
5	23	22	21	12	32	30	3	54	33
6	33	89	53	65	45	42	2	10	23
7	30	33	55	42	87	30	2	10	30
8	36	56	32	3	65	86	54	23	30
9	23	3	78	63	68	68	54	12	39
10	98	65	63	32	45	75	86	63	21
11	53	65	86	96	63	32	12	45	25
12	33	22	42	21	3	35	63	32	54
13	10	30	53	65	43	20	32	45	65
14	32	35	63	65	66	33	53	63	32
15	12	30	56	22	30	56	42	30	12
16	56	33	89	65	78	99	63	30	24
17	53	30	36	65	22	33	22	54	33
18	32	30	36	65	63	32	30	36	65
19	30	36	65	66	33	22	12	30	32
20	98	65	63	65	45	12	2	36	30
21	00	22	11	42	53	63	32	53	32
22	30	35	63	66	33	22	56	52	21

coefficients defining the orthonormal variables associated with the within-subject effects B, C, and BC. Let $\mathbf{S(X)}$ be the unbiased estimate of the covariance matrix of the nine repeated measures. Then the sample covariance matrices of the orthonormal variables associated with B, C, and BC are, respectively,

$$\mathbf{S}(\mathbf{Y}_B) = \mathbf{M}_B\mathbf{S}(\mathbf{X})\mathbf{M}'_B,\ \mathbf{S}(\mathbf{Y}_C) = \mathbf{M}_C\mathbf{S}(\mathbf{X})\mathbf{M}'_C,$$

and $\mathbf{S}(\mathbf{Y}_{BC}) = \mathbf{M}_{BC}\mathbf{S}(\mathbf{X})\mathbf{M}'_{BC}$. The corresponding sphericity criteria are $W_B = .9921$, $W_C = .9923$, and $W_{BC} = .5055$. At the .05 level of significance, the critical values are .7411, .7411, and .4173, respectively. Since large values of W support the sphericity assumption, the data indicate that the validity conditions for the B, C, and BC traditional F tests are tenable.

Example 2

The 22 subjects of Table 2 are now assigned to two levels of the independent factor A. The first level consists of the first 10 subjects, and the second level has 12 subjects. The orthonormal variables associated with the clusters of within-subject effects (B, AB), (C, AC), and (BC, ABC) are defined by the matrices $\mathbf{M}_B$, $\mathbf{M}_C$, and $\mathbf{M}_{BC}$ as in Numerical Example 1. The Box criteria for equality of covariance matrices are, respectively, $M_B = 2.4626$, $M_C = 1.7064$, and $M_{BC} = 10.2912$. The critical values at the .025 level of significance are $\rho_B\chi^2(3) = .8902 \times 9.3484 = 8.3219$ for M_B and M_C, and $\rho_{BC}\chi^2(4) = .7821 \times 11.1433 = 8.7152$ for M_{BC}. The Box criteria are thus small enough to support the presumption of equal covariance matrices across the two levels of A for each of the three sets of orthonormal variables $\mathbf{Y}_B$, $\mathbf{Y}_C$, and $\mathbf{Y}_{BC}$.

The preliminary testing may now proceed to the hypothesis of sphericity for each common covariance matrix. The Mauchly criteria for the clusters (B, AB), (C, AC), and (BC, ABC) are .9912, .9780, and .4846, respectively. It may be noted that under the assumption of equality of the covariance matrices, each pooled sample covariance matrix with 20 degrees of freedom may be considered

to be based on a random sample of 21 subjects. By entering the value of $N = 21$ in the tables of Nagarsenker and Pillai (1972), the interpolated critical values at the .025 level are .6776 for the first two clusters and .3542 for the last cluster. Hence, the assumption of sphericity holds for each set of orthonormal variables.

In summary, the data analysis (at the joint .05 level of significance) indicates that assuming normality, the validity conditions for the traditional within-subject F tests hold for the data of this numerical example.

Effect of Nonnormality on the Mauchly Test

As indicated previously, the Mauchly test is appropriate in all cases to check the validity of the traditional testing procedure. Being by construction a *LR* criterion test, W may be suspected to be oversensitive to departure from normality. This is known to occur for various *LR* tests of variance homogeneity such as the Bartlett test and two of its competitors, the Hartley $F_{\max}$ and the Cochran criteria (Scheffé, 1959; also see Games, Winkler, & Probert, 1972, for a list of references). On the other hand, there are indications that least squares procedures (which are equivalent to *LR* methods in some situations) are fairly robust when the populations involved have light tails (Hogg, 1974; Tukey & McLaughlin, 1963, p. 332). As noted by Hogg, there are many practical situations in which distributions are inherently light tailed. This is particularly true in the social sciences if the measuring instrument has modest floor and ceiling effects. However, data may contain outliers that tend to shift the density toward the tails, thus creating situations in which tails are heavier than that of the normal distribution. Since the F test for means is fairly insensitive to the shape of the parent distributions and since the W criterion may be used to determine the tenability of the F test, it is desirable to explore the effect of nonnormality on W. The remainder of this article focuses on the Type I error rate associated with the W criterion under several instances involving light-tailed distributions, heavy-tailed distributions, and mixtures of normal distributions. These results were obtained using the technique of computer simulation for situations in which the p components of the vector $\mathbf{X}$ were independent and had the same distribution. For normal distributions, the standardized fourth moment (β_2, a measure of kurtosis) = 3. Light-tailed distributions correspond to $\beta_2 < 3$ and heavy-tailed distributions to $\beta_2 > 3$.

Selection of Common Distributions (Component Distributions)

Eight component distributions were selected to represent a variety of departures from normality. The three light-tailed distributions with bounded range that were chosen were the uniform distribution on the $0 - 1$ interval ($\beta_2 = 1.8$), the convolution of two such uniform distributions (a triangular distribution with $\beta_2 = 2.4$), and the convolution of three such uniform distributions (a trapezoidal distribution with $\beta_2 = 2.6$). They are subsequently denoted as U1, U2, and U3, respectively.

Five heavy-tailed component distributions were also selected. The first represents the distribution of the product of two mutually independent variables, one having the $0 - 1$ uniform distribution and the other being normally distributed with zero mean and unit variance [i.e., $N(0, 1)$]. This distribution ($\beta_2 = 5.4$) is labeled UN. The second heavy-tailed distribution was chosen to be the Laplace (or double exponential) distribution for which $\beta_2 = 6$. The remaining three heavy-tailed distributions were mixtures of two normal distributions, each with mean zero and variances of 1 and 9, respectively [i.e., $N(0, 1)$ and $N(0, 9)$]. The mixtures were denoted as $(1 - \lambda)\ N(0, 1) + \lambda N(0, 9)$, and the mixing proportions λ were set at 5% ($\beta_2 = 7.653$), 10% ($\beta_2 = 8.333$), and 20% ($\beta_2 = 7.544$).

Simulation Process

A computer program was written to simulate n independent vectors, each having p components drawn independently from the same component distribution. A sample co-

variance matrix S based on these n vectors was then obtained, and the criterion W was derived from S. The appropriate critical value for W was retrieved from tables in Nagarsenker and Pillai (1973). The empirical proportion of Type I errors was obtained by dividing the number of times that W exceeded the given value by the number of data sets simulated. Five thousand replications were made for each combination of n, p, and component distribution to estimate the true Type I error at nominal alpha values of 10%, 5%, 2.5%, and 1%. To check the accuracy of the simulation process, initial computer runs were made using the normal distribution. As may be seen in Table 3, the discrepancies between the empirical Type I errors and their respective true values for the normal case are within 2.3 standard errors for proportions.

Results

Though the simulation was conducted for $p = 2$, 3, 4, and 5, combined with several levels of sample size n, only the data for $p =$ 5 are reported in Table 3. The patterns displayed by the empirical Type I error for $p =$ 2, 3, and 4 are virtually identical to those observed for the case $p = 5$.

The following trends may be deduced from Table 3.

1. The Mauchly W criterion tends to err on the conservative side for light-tailed component distributions, and the discrepancy between the empirical Type I error and nominal alpha is greater for large samples.

2. With heavy-tailed component distributions, the empirical Type I error rates are much larger than the corresponding nominal values. As in the previous case, the differences become more visible as the sample size increases.

3. In all situations under consideration, the ratio of the empirical Type I error to the posted alpha deviates further from unity at smaller alpha values. This trend has also been noticed in most studies regarding the robustness of the F test under nonnormality and/or variance heterogeneity (Collier et al., 1967; Glass, Peckham, & Saunders, 1972; Huynh, 1969).

Discussion of Simulation Study

In the simulation study, the behavior of Mauchly's sphericity criterion W has been documented for a number of nonnormality conditions. For light-tailed distributions, W errs on the conservative side in terms of Type I error. This is partially because W is an increasing function of the variability of the eigenvalues of the sample covariance matrix S. A light-tailed distribution usually is more dense near the mean than is true of a normal distribution. Thus, the eigenvalues calculated from a covariance matrix based on a light-tailed distribution would be expected to display less heterogeneity than those calculated from a covariance matrix based on a normal distribution with the same variance. Hence the critical values under normality for W will be somewhat larger than are appropriate for values of W calculated from light-tailed distributions.

The simulation previously reported clearly indicates that W does not behave badly as long as the component distribution has a tail lighter than that of the normal distribution. Fortunately, most data in the social sciences are obtained from measuring instruments with a limited score range. The corresponding distributions are hence likely to be light-tailed, though probably not as extreme as the uniform distribution. It would be expected that W, along with the normal distribution critical values, should do a reasonable job in terms of Type I error in checking the sphericity assumptions in designs involving repeated measures.

The study also points out the fallibility of the Mauchly criterion in the case of heavy-tailed component distributions. The authors are not aware of a method of overcoming this deficiency, although nonparametric or robust procedures could conceivably be developed. Though W may be relied on in preliminary testing for the validity of the traditional F tests, it may simply be skipped in many instances. In previous articles on approximate tests for repeated measures designs, Huynh and Feldt (1976) and Huynh (1978) have shown that an adjustment in the degrees of freedom would be sufficient to account for most examples of departure of the covariance

Table 3
Empirical Percentages of Type I Error for Mauchly's Sphericity Criterion W Under Selected Situations

Component distribution	Alpha percentages															
	10[a]	5[a]	2.5[a]	1[a]	10[b]	5[b]	2.5[b]	1[b]	10[c]	5[c]	2.5[c]	1[c]	10[d]	5[d]	2.5[d]	1[d]
Normal	9.6	5.0	2.7	.8	9.5	4.7	2.2	.8	10.1	4.8	2.6	1.2	10.9	5.7	2.8	1.1
U1	6.1	2.8	1.1	.5	5.4	2.6	1.3	.6	5.0	2.4	1.2	.6	4.5	2.1	.8	.3
U2	8.3	4.1	1.7	.8	7.1	3.4	1.8	.8	7.1	3.4	1.6	.6	6.2	2.9	1.2	.4
U3	8.6	4.2	2.0	.7	8.2	3.7	1.9	.7	7.9	3.9	1.7	.7	8.0	3.6	1.5	.5
UN	23.8	14.7	9.2	5.1	26.9	17.8	11.3	5.8	29.1	19.4	12.8	7.3	30.0	20.0	13.3	7.3
Laplace	22.6	13.3	8.0	3.9	28.2	18.5	12.4	6.9	30.9	20.6	13.9	7.8	32.7	22.7	15.3	9.0
Mixed (λ = 5%)	21.9	13.8	8.5	4.7	26.6	18.3	12.3	7.4	32.0	22.7	16.5	10.5	34.2	24.5	17.4	11.3
Mixed (λ = 10%)	26.8	17.9	11.8	6.5	35.3	24.8	17.6	10.9	39.9	29.0	21.3	13.4	43.7	33.0	25.0	17.3
Mixed (λ = 20%)	31.7	20.8	13.6	7.6	39.1	26.9	18.6	12.0	41.8	30.2	22.7	14.2	46.4	34.7	25.9	17.6

Note. U1 = uniform distribution on the 0 − 1 interval; U2 = the convolution of two such uniform distributions; U3 = the convolution of three such uniform distributions; UN = Uniform × Normal Distribution; Laplace = double exponential distribution.
[a] $n = 10$.
[b] $n = 15$.
[c] $n = 20$.
[d] $n = 26$.

matrix from sphericity. If heavy-tailed component distributions are suspected, simply discard the Mauchly sphericity criterion and conduct an approximate test. In most instances the true probability of the Type I error will not be far from the posted nominal alpha. For readers familiar with multivariate analysis, of course, alternative testing procedures such as Hotelling's T and many others based on the union-intersection principle are available (Morrison, 1976).

Conclusions

Details regarding the necessary and sufficient conditions for traditional tests of the within-subject mean square ratios in repeated measures designs are presented in this article. It is shown that these conditions are based on the orthogonal normalized variables associated with each cluster of within-subject mean square ratios based on the same error term. They are far more general than the assumptions of equality of covariance matrices and compound symmetry of the common covariance matrix for all repeated measures. It is shown that appropriate preliminary testing for the mentioned conditions may be carried out within normal distributions via the Box modified likelihood ratio test and the Mauchly sphericity criterion. Both tests are to be conducted on suitably chosen sets of orthonormal variables.

Furthermore, the behavior of the Mauchly criterion for nonnormal data was investigated. It was shown to provide a conservative testing procedure for light-tailed component distributions and to produce more than the nominal percentage of Type I errors for heavy-tailed distributions. Since many nonnormal distributions that occur in the social sciences are light-tailed, the W criterion should be useful for assessing the validity of the application of the traditional tests of repeated measures data.

The other test, the Box test, has also been suspected to be sensitive to nonnormality. Further studies are needed to determine the seriousness of this problem and possibly to identify procedures to overcome it.

Reference Note

1. Nie, N. H., & Hull, H. *Statistical package for the social sciences batch release 7.0 update manual.* Mimeographed manual, March 1977. (Available from Norman H. Nie, Suite 3300, 444 North Michigan Avenue, Chicago, Illinois 60611.)

References

Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. Estimates of test sizes for several test procedures based on conventional variance ratios in the repeated measures design. *Psychometrika,* 1967, *32,* 339–353.

Consul, P. C. On the exact distribution of the criterion W for testing sphericity in a p-variate normal distribution. *Annals of Mathematical Statistics,* 1967, *38,* 1170–1174.

Consul, P. C. The exact distribution of likelihood criteria for different hypotheses. *Multivariate analysis* (Vol. 2). New York: Academic Press, 1969.

Davidson, M. L. Univariate versus multivariate tests in repeated measures experiments. *Psychological Bulletin,* 1972, *77,* 446–452.

Ferguson, G. A. *Statistical analysis in psychology and education* (4th ed.). New York: McGraw-Hill, 1976.

Games, P. A., Winkler, H. B., & Probert, D. A. Robust tests for homogeneity of variance. *Educational and Psychological Measurement,* 1972, *32,* 887–909.

Glass, G. V, Peckham, P. D., & Saunders, P. Consequences of failure to meet assumptions underlying the analysis of variance and covariance. *Review of Educational Research,* 1972, *42,* 237–288.

Hogg, R. V. Adaptive robust procedures: A partial review and some suggestions for further applications and theory. *Journal of the American Statistical Association,* 1974, *69,* 909–927.

Huynh, H. Effect of heterogeneity of covariance on the level of significance of certain proposed tests for the treatment by subject and Type I designs (Doctoral dissertation, University of Iowa, 1969). *Dissertation Abstracts International,* 1969, *30,* 42A. (University Microfilms No. 69–13,153)

Huynh, H. Some approximate tests for repeated measurement designs. *Psychometrika,* 1978, *43,* 161–175.

Huynh, H., & Feldt, L. S. Conditions under which mean square ratios in repeated measurement designs have exact F-distributions. *Journal of the American Statistical Association,* 1970, *65,* 1582–1589.

Huynh, H., & Feldt, L. S. Estimation of the Box correction for degrees of freedom from sample data in the randomized block and splitplot designs. *Journal of Educational Statistics,* 1976, *1,* 69–82.

Keppel, G. *Design and analysis: Researcher's handbook.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Kirk, R. E. *Experimental design: Procedures for the behavioral sciences.* Monterey, Calif.: Brooks/Cole, 1968.

Korin, B. P. On testing the equality of k covariance matrices. *Biometrika*, 1969, *56*, 216–218.

Mathai, A. M., & Rathie, P. N. The exact distribution for the sphericity test. *Journal of Statistical Research*, 1970, *4*, 140–159.

Mauchly, J. W. Significance test for sphericity of n-variate normal populations. *Annals of Mathematical Statistics*, 1940, *11*, 204–209.

Mendoza, J. L., Toothaker, L. E., & Cain, B. R. Necessary and sufficient conditions for F ratios in the $L \times J \times K$ factorial design with two repeated factors. *Journal of the American Statistical Association*, 1976, *71*, 992–993.

Morrison, D. F. *Multivariate statistical methods* (2nd ed.). New York: McGraw-Hill, 1976.

Nagarsenker, B. N., & Pillai, K. C. S. *The distribution of the sphericity test criterion* (ARL 72–0154). Wright-Patterson Air Force Base, Ohio: Aerospace Research Laboratory, November 1972. (NTIS No. AD-754 232)

Nagarsenker, B. N., & Pillai, K. C. S. The distribution of the sphericity test criterion. *Journal of Multivariate Analysis*, 1973, *3*, 226–235.

Pearson, E. S., & Hartley, H. O. *Biometrika tables for statisticians* (Vol. 2). Cambridge, England: Cambridge University Press, 1972.

Pillai, K. C. S., & Nagarsenker, B. N. On the distribution of the sphericity test criterion in classical and complex normal populations having unknown covariance matrices. *Annals of Mathematical Statistics*, 1971, *42*, 764–767.

Poor, D. S. D. Analysis of variance for repeated measures design: Two approaches. *Psychological Bulletin*, 1973, *80*, 204–209.

Rouanet, H., & Lepine, D. Comparison between treatments in a repeated measures design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, 1970, *23*, 147–163.

Scheffé, H. *The analysis of variance*. New York: Wiley, 1959.

Timm, N. H. *Multivariate analysis with application in education and psychology*. Monterey, Calif.: Brooks/Cole, 1975.

Tukey, J. W., & McLaughlin, D. H. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhyā*, 1963, *A25*, 331–352.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Received March 13, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 974–977

Large Sample Variance of Kappa in the Case of Different Sets of Raters

Joseph L. Fleiss
School of Public Health, Columbia University, and Biometrics Research Department, New York State Psychiatric Institute, New York

John C. M. Nee
Biometrics Research Department, New York State Psychiatric Institute, New York

J. Richard Landis
Department of Biostatistics, School of Public Health, University of Michigan

Published formulas for the large sample variance of the kappa statistic that are appropriate for the case of different sets of raters for different subjects, when each set of raters is selected at random from a larger pool of available raters, are determined to be incorrect. New formulas are derived and checked by Monte Carlo simulation. Kappa is shown to be identical, except for terms that go to zero as the number of subjects increases, to the intraclass correlation coefficient resulting from applying a one-way analysis of variance to the data.

Many human endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for kappa is a third.

Cohen (1960, 1968) presented kappa and weighted kappa as chance-corrected measures of agreement between two raters, each of whom independently classifies each of a sample of subjects into one of k mutually exclusive and exhaustive categories. The standard error formulas that he presented as well as formulas published by Everitt (1968) were shown by Fleiss, Cohen, and Everitt (1969) to be incorrect. The formulas presented in the latter article have been confirmed analytically by Landis and Koch (1977a) and have been confirmed by means of Monte Carlo simulation by Cicchetti and Fleiss (1977) and Fleiss and Cicchetti (1978).

Fleiss (1971) extended kappa to the case in which each of a sample of subjects is rated on a nominal scale by the same number of raters but in which the raters rating one subject are not necessarily the same as those rating another. The standard error formulas presented in that article are incorrect. Landis and Koch (1977b) derived similar statistics for the case of possibly varying numbers of ratings per subject by applying a one-way analysis of variance model to the data. The method they suggested for calculating standard errors yields results appropriate for the nonnull case but overestimates the standard error appropriate for testing the null hypothesis that the parameter is zero.

In this article, formulas for the standard error of kappa in the case of different sets of equal numbers of raters that are valid when the number of subjects is large and the null hypothesis is true are derived. The results of some Monte Carlo simulations confirm that the formulas are correct.

Notation

The notation here is the same as in Fleiss (1971). Let N represent the total number of subjects, n the number of ratings per subject, and k the number of categories into which assignments are made. Let the subscript i, where $i = 1, \ldots, N$, represent the subjects, and the subscript j, where $j = 1, \ldots, k$, represent the categories of the scale.

Requests for reprints should be sent to Joseph L. Fleiss, Division of Biostatistics, Columbia University School of Public Health, 600 West 168th Street, New York, New York 10032.

Define n_{ij} to be the number of raters who assigned the ith subject to the jth category, and define

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}. \tag{1}$$

The quantity p_j is the proportion of all assignments that were to the jth category. Since $\Sigma_j n_{ij} = n$, $\Sigma_j p_j = 1$.

A motivation of the following formulas, which are simplifications of the expressions presented there, can be found in Fleiss (1971). The measure of the extent of agreement beyond chance in assigning subjects to category j ($j = 1, \ldots, k$) is

$$\kappa_j = 1 - \frac{\sum_i n_{ij}(n - n_{ij})}{Nn(n-1)p_j q_j}, \tag{2}$$

where $q_j = 1 - p_j$. If there is perfect agreement in the assignments to category j (i.e., if each $n_{ij} = 0$ or n), then $\kappa_j = 1$. If, on the other hand, the n_{ij}s vary as binomial random variables with parameters n and p_j, then the expected value of κ_j is 0. The minimum value of κ_j is $-1/(n-1)$.

The overall measure of agreement beyond chance is a weighted average of the κ_js,

$$\kappa = \frac{\sum p_j q_j \kappa_j}{\sum p_j q_j} = 1 - \frac{Nn^2 - \sum \sum n_{ij}^2}{Nn(n-1) \sum p_j q_j}. \tag{3}$$

The overall measure also varies from a minimum of $-1/(n-1)$ for poorer than chance agreement through 0 for just chance agreement to unity for perfect agreement.

Large Sample Standard Errors

The error committed by Fleiss in his 1971 article was to ignore the fact that the denominators of κ and κ_j are subject to the same order of random variation as are their numerators. The results below are derived by taking proper account of the variation in both the numerators and denominators.

It is convenient to define

$$s_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}^2, \tag{4}$$

so that kappa (see Equation 3) may be reexpressed as

$$\kappa = 1 - \frac{n - \sum s_j}{(n-1)(1 - \sum p_j^2)}. \tag{5}$$

Consider the hypothesis that the ratings are purely random in the sense that for each subject, the frequencies $n_{i1}, n_{i2}, \ldots, n_{ik}$ are a set of multinominal frequencies with parameters n and $(P_1, P_2, \ldots, P_k)$, where $\Sigma P_j = 1$. Using known results about moments of powers and products of multinomial variables (the moments given by Fleiss, 1971, in Equations 12–15 are correct, except that the term n_{ij} in Equation 14 should be squared), it may be checked that the covariance matrix Σ of $(p_1, s_1, p_2, s_2, \ldots, p_k, s_k)'$ is given by the following expression:

$$\Sigma = \begin{Vmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1k} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2k} \\ \vdots & \vdots & & \vdots \\ \Sigma_{k1} & \Sigma_{k2} & \cdots & \Sigma_{kk} \end{Vmatrix}, \tag{6}$$

where, letting $Q_j = 1 - P_j$ and

$$F_j = 1 + 2(n-1)P_j, \tag{7}$$

$$\Sigma_{ij} = \frac{-P_i P_j}{Nn} \begin{bmatrix} 1 & F_j \\ F_i & F_i F_j - 2(n-1)P_i P_j \end{bmatrix}, \tag{8}$$

for $i \neq j$, and

$$\Sigma_{jj} = \frac{P_j Q_j}{Nn} \begin{bmatrix} 1 & F_j \\ F_j & F_j + (F_j - 1)Q_j \end{bmatrix}. \tag{9}$$

Under the hypothesis of randomness, the expected value of p_j is P_j, and that of s_j is $P_j[1 + (n-1)P_j]$. The vector $\mathbf{v}$ of partial derivatives of κ with respect to each of its components $(p_1, s_1, \ldots, p_k, s_k)$, evaluated at the parameter values, is

$$\mathbf{v} = \begin{Vmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_k \end{Vmatrix}, \tag{10}$$

where

$$\mathbf{v}_i = \frac{1}{(n-1)\sum P_j Q_j} \begin{bmatrix} -(F_i - 1) \\ 1 \end{bmatrix}. \tag{11}$$

According to standard large sample theory (Rao, 1973), the approximate variance of κ when N is large may be found by replacing

the unknown parameters by their sample estimates in $\mathbf{v}'\Sigma\,\mathbf{v}$. It may be checked by straightforward algebra, then, that the estimated large sample variance of κ is

$$\mathrm{Var}(\kappa) = \frac{2}{Nn(n-1)(\sum p_j q_j)^2} \times [(\sum p_j q_j)^2 - \sum p_j q_j (q_j - p_j)]. \quad (12)$$

The variance of κ_j may be found either by applying the same kind of algebra as above or, more directly, by specializing Equation 12 to the binomial case where $k = 2$ by considering only ratings into or not into category j. In either case, the large sample variance of κ_j is

$$\mathrm{Var}(\kappa_j) = \frac{2}{Nn(n-1)}, \quad (13)$$

independent of the proportions.

For large values of N and under the hypothesis of randomness, κ and the separate κ_js are approximately normally distributed with variances given by Equations 12 and 13. This result follows from the multivariate central limit theorem (Rao, 1973) applied to the average values $\{p_j\}$ and $\{s_j\}$. The kappas are slightly negatively biased, however. This follows from the identity

$$\kappa_j = \frac{\chi^2_j - N}{N(n-1)}, \quad (14)$$

where

$$\chi^2_j = \frac{\sum_{i=1}^{N} (n_{ij} - np_j)^2}{np_j q_j}, \quad (15)$$

the chi-square statistic for testing the homogeneity of N binomial samples. Because the expected value of χ^2_j under the hypothesis is approximately equal to $N - 1$, that of κ_j (and of κ) under the hypothesis is approximately equal to $-1/N(n-1)$.

Thus, when N is large, the statistical significance of κ_j may be tested by referring the quantity

$$z_j = \left[\kappa_j + \frac{1}{N(n-1)}\right]\sqrt{\frac{Nn(n-1)}{2}} \quad (16)$$

to the standard normal distribution. The statistical significance of κ may be tested by referring the quantity

$$z = \left[\kappa + \frac{1}{N(n-1)}\right](\sum p_j q_j) \times \sqrt{\frac{Nn(n-1)}{2[(\sum p_j q_j)^2 - \sum p_j q_j (q_j - p_j)]}} \quad (17)$$

to the standard normal distribution.

The empirical distributions of z and of $\{z_j\}$ were obtained by Monte Carlo simulation for a number of combinations of parameter values. The empirical standard deviations were all close to the theoretical value of unity, and if N is at least 25 or 30, then testing the statistical significance of κ and $\{\kappa_j\}$ by referring the values of z and $\{z_j\}$ to the standard normal distribution seems safe.

The incorrect formulas given in Equations 16 and 23 of Fleiss (1971) overestimate the variance, as do the formulas proposed by Landis and Koch (1977b) for the nonnull case. The use of these formulas, therefore, together with a failure to take account of the negative bias in the kappas, leads to conservative tests of significance.

Kappa As an Approximate Intraclass Correlation Coefficient

Landis and Koch (1977b) approach the problem of measuring the degree of agreement on the jth category by applying the algebra of a one-way analysis of variance to the data resulting from coding assignments to the jth category as 1 and assignments to another category as 0. The mean square within subjects is equal to

$$WMS_j = \frac{1}{Nn(n-1)} \sum_i n_{ij}(n - n_{ij}), \quad (18)$$

and the mean square between subjects is equal to

$$BMS_j = \frac{1}{(N-1)n} \sum_i (n_{ij} - np_j)^2. \quad (19)$$

They take as the measure of agreement on the jth category the sample intraclass correlation coefficient

$$r_{jj} = \frac{BMS_j - WMS_j}{BMS_j + (n-1)WMS_j}, \quad (20)$$

and as the overall measure of agreement a weighted average of these coefficients,

$$\bar{r} = \frac{\sum p_j q_j r_{jj}}{\sum p_j q_j}. \tag{21}$$

It is easily checked that

$$r_{jj} = \frac{1 + \kappa_j(Nn - 1)}{(Nn - n + 1) + \kappa_j(n - 1)}, \tag{22}$$

which is always larger than κ_j. If N is large, however, r_{jj} and κ_j are virtually equal. In fact, if BMS_j is redefined to have N instead of $N - 1$ in its denominator, then r_{jj} and κ_j are identical.

A major contribution by Landis and Koch (1977b) to the measurement of agreement was their consideration of the case of varying numbers of ratings per subject. In addition, they indicated how large sample variances appropriate to the nonnull case could be calculated.

References

Cicchetti, D. V., & Fleiss, J. L. Comparison of the null distributions of weighted kappa and the C ordinal statistic. *Applied Psychological Measurement*, 1977, *1*, 195–201.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, *20*, 37–46.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, *70*, 213–220.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 97–103.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 1971, *76*, 378–382.

Fleiss, J. L., & Cicchetti, D. V. Inference about weighted kappa in the non-null case. *Applied Psychological Measurement*, 1978, *2*, 113–117.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323–327.

Landis, J. R., & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 1977, *33*, 159–174. (a)

Landis, J. R., & Koch, G. G. A one-way components of variance model for categorical data. *Biometrics*, 1977, *33*, 671–679. (b)

Rao, C. R. *Linear statistical inference and its applications* (2nd ed.). New York: Wiley, 1973.

Received March 27, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 978–984

Tests for Homogeneity of Variance in Factorial Designs

Paul A. Games
Pennsylvania State University

Harvey J. Keselman and Jennifer J. Clinch
University of Manitoba, Winnipeg, Canada

The variance portion of Games's three-factor model of inference on independent groups is extended. Six procedures that convert tests of spread into tests of location are reviewed and explored in a Monte Carlo study of how to test variances in a factorial design. The statistic $ln\ s^2$ is shown to be a slight improvement over Overall and Woodward's procedure. The dependence of these two tests on the normality condition is illustrated. Four robust alternatives of somewhat lower power are contrasted. The jackknife test is the most powerful and is only slightly sensitive to leptokurtosis if the ns are equal. The Brown-Forsythe median test is acceptable but it uses average deviations rather than variances. The Box-Scheffé test is always robust. No single test is ideal. A two-stage process is recommended.

Behavioral investigators often attend only to central tendency, even when interesting trends in variability exist in the data. In education, it is desirable to reduce the variance when working with hierarchial tasks. Students will show little divergence on entry behaviors for the next task in the hierarchy, if we have been able to keep the variance small in the first task. Skinner (1958) suggested reduced variances as a desirable consequence from programmed learning, and Block (1973) recognized in his discussions of mastery learning that this outcome was desirable. Unfortunately, most investigators have failed to attend to variances except as a "nasty" assumption with respect to the analysis of variance (ANOVA) on means. Two exceptions to this trend are Birch and Lefford (1967) and Johnson and Baker (1973).

This neglect of variances is probably related to the fact that typical statistics texts teach only the classical omnibus tests on homogeneity of variance that (a) ignore the logical structure of factorial designs and (b) are extremely sensitive (nonrobust) to violations of the normality assumption. The present article presents six tests that can be used in factorial designs of independent groups in a fashion similar to that of the familiar tests on means. Three of the six tests prove to be robust.

Tests Requiring Normality

Overall and Woodward (1974) proposed the Z-variance test, a clever extension and application of the Fisher and Yates (1963) z-score transformation for chi-square statistics. There are problems of clarity in the formulation of the statistic and in the article, however. Given K samples of sizes n_k from independent populations, Overall and Woodward define

$$Z_k = \sqrt{\frac{c_k(n_k - 1)s_k^2}{MS_w}} - \sqrt{c_k(n_k - 1) - 1}, \quad (1)$$

where $c_k = 2 + 1/n_k$ and MS_w is mean square within cells. Overall and Woodward then define

$$F_{(K-1,\infty)} = \sum^{K} Z_k^2/(K - 1). \quad (2)$$

They note that $(K-1)F_{(K-1,\infty)} = \chi^2_{(K-1)}$ and refer to Marascuilo's (1966) statistic $U' = \Sigma[(\hat{\theta}_k - \hat{\theta}_o)^2/\mathrm{var}(\hat{\theta}_k)]$ that is distributed as $\chi^2_{(K-1)}$. $\hat{\theta}_k$ is a normally distributed unbiased estimate of the parameter θ_k, θ_o is the common value of θ_k under the hypothesis of equality,

Requests for reprints should be sent to Paul A. Games, Division of Counseling and Educational Psychology, Pennsylvania State University, 403D Carpenter Building, University Park, Pennsylvania 16802.

and $\hat{\theta}_o$ is an estimate of θ_o. Assuming $\text{var}(Z_k) = 1$ and $\theta_o = 0$, Overall and Woodward derive Equation 2. However, they later recommend, "The z-transformed sample variances can be analyzed by ANOVA for factorial design with one observation per cell" (1974, p. 313). The standard ANOVA will compute deviations squared from the grand mean, in this case, $\bar{Z}$. This procedure is used in their illustrative example, hence they have used a different procedure than in their definition in Equation 2. This problem becomes particularly severe in unequal n factorial designs, in which there are different ways that the row and column marginal means and the grand mean can be defined. Their procedure is analogous to the unweighted means modification of ANOVA. Carlson and Timm (1974) and Applebaum and Cramer (1974) illustrate the complexities of the more desirable least squares analysis of nonorthogonal designs. To avoid these complexities, in the present article we stick with the equal n case.

Bartlett and Kendall (1946) proposed an alternative transformation and illustrated its usage in a two-factor design. They propose using $v_k = ln\, s_k^2$ as the transformed value to compute on each cell and to enter into an ANOVA. (Either $ln\, s^2 = \log_e s^2$ or $\log_{10} s^2$ could be used, since one is a linear transform of the other. The $ln\, s^2$ usage simplifies later results.)

Overall and Woodward (1974) discard this statistic because Bartlett and Kendall "themselves conclude that the ANOVA of log variances is inferior to the more frequently used Bartlett's test (1937) in situations where the two can be compared" (p. 311). However, the inferiority of v_k was demonstrated when equal ns below 12 were used, values of n that were smaller than those used by Overall and Woodward or Levy (1975) in their Monte Carlo studies of Z_k. On such small samples, we also can expect the Z-variance test to be inferior to the traditional one-factor Bartlett test. Unfortunately, neither the Overall and Woodward nor the Levy studies contrasted the two.

On any given set of data, MS_w will be a constant, so the definition can be rewritten as

$$Z_k = \sqrt{\frac{c(n-1)}{MS_w}}\,\sqrt{s_k^2} - \sqrt{c(n-1) - 1} = a\sqrt{s_k^2} - b. \quad (3)$$

Thus Z_k is a linear transform of s_k, the square root of s_k^2. The square root transformation is appropriate for stabilizing the variance of the statistic, if the mean of the statistic is proportional to the variance of the statistic; but the ln transformation is appropriate if the mean is proportional to the standard deviation (Bartlett, 1947). The latter is the case with s^2. $E(s^2) = \sigma^2$, and $\sigma_{s^2} = \sigma^2[(2/n - 1) + (\gamma_2/n)]^{\frac{1}{2}}$, where γ_2 is the kurtosis index of the population. Thus, for the equal n case, use of $v_k = ln\, s_k^2$ will stabilize the variance of the ANOVA entries better than use of Z_k. When H_o is true, and the s_k^2 statistics are reasonably closely clustered, the results of the two transformations will be similar, but as the s_k^2 statistics become more divergent, the difference increases. Thus the effect on power should be greater than the effect on the familywise risk of Type I errors (FWI; Games, 1971).

The disadvantage of both above formulations is that they are exceedingly sensitive to violations of the normality assumption, as are all other classical tests on variances (Box, 1953). Scheffé (1959) shows that $E(v) \approx ln\, \sigma^2$, and $\text{var}(v) \approx (2/(n-1)) + (\gamma_2/n)$. If a good estimate of the common γ_2 is available, it could be used. Box and Anderson (1955) build robust tests for variances by use of an estimate of γ_2. Such estimates require large ns for stability, however, and add great complexity to the computation, while the statistics that Box and Anderson proposed do not readily extend to multifactor cases.

The Scheffé (1959) formulation makes it obvious why the v_k test is so sensitive to nonnormality. By assuming normality, γ_2 is set to 0 to yield $MS_e = 2/(n-1)$. If the populations are platykurtic, then $\gamma_2 < 0$, and this theoretical MS_e is too large, which results in a conservative test. With leptokurtic populations, $\gamma_2 > 0$, so use of the theoretical MS_e results in an excess of Type I errors. Games, Winkler, and Probert (1972) showed that only the latter condition is a problem. Platykurtic populations yield greater power on the classical tests than do normal populations, despite the conservatism at the null hypothesis point. Why object to a reduced risk of Type I error if there is no corresponding power loss?

Tests Robust to Nonnormality

Both Box (1953) and Scheffé (1959) suggest breaking each cell into random subsamples and computing $ln\, s^2$ on each subsample. If each cell has n cases, then I subsamples of m cases each ($n = mI$ when possible) are determined, and values of $v_{ijk} = ln\, s_{ijk}^2$, $i = 1, \ldots I$, are computed in each cell. These values are used as input to an ANOVA with m observations per cell rather than with one observation per cell as in the prior tests. Now MS_w may be computed and is an unbiased estimate of var(v), whether Y is normally distributed or not. This technique has been labeled the Scheffé test (Winer, 1971), the Box-Scheffé test (Levy, 1975), and the Bartlett and Kendall test (Games et al., 1972; Gartside, 1972). We use the Box-Scheffé label here to avoid confusion with the prior Bartlett and Kendall suggestion of using a single value of $ln\, s^2$ per cell.

Games et al. (1972), Games (1975), and Levy (1975) pointed out that the Box-Scheffé test can be used on multifactor designs and is robust to violations of the normality assumption. Levy compared the power of the Overall and Woodward (1974) Z-variance test and the Box-Scheffé test for single-factor designs with equal n and $K = 3$, when Y is normally distributed. Levy concluded, "For all sample sizes, one can plainly see that the Z-variance test is vastly superior to the Box-Scheffé procedure with respect to power" (p. 521). Levy's conclusion is due, however, to an exceptionally poor choice of subsample size for the computation of the Box-Scheffé test. Levy used $m = 2$ for ease of computation in his Monte Carlo study. However, Gartside (1972) and Games et al. demonstrated that the use of subsamples of only two cases produces power far lower than use of intermediate subsample sizes. Martin and Games (Note 1) further investigated the desirable subsample size and concluded that the use of $m \approx (n)^{\frac{1}{2}}$ (or the nearest whole divisor of n, if any) results in optimum power in the Box-Scheffé test. Games et al. report a procedure for a rough estimation of the power of the Box-Scheffé test. When appropriate values of $m = 3$ for $n = 12$, $m = 5$ for $n = 26$, and $m = 6$ for $n = 40$ are used, the estimated powers are uniformly higher than the powers that Levy reports. In the trade-off between robustness and power, the cost of the Box-Scheffé, properly used, is far less than implied by Levy's results or conclusions.

The major disadvantage of the Box-Scheffé test is that the use of random subsamples makes it possible for different data analysts to obtain different outcomes from the same data. This is unlikely with clear-cut data but might be a problem with borderline data. Scheffé (1970) reports that some users randomized and rerandomized until they obtained the results they wanted on a related test using subsample randomization. Brown and Forsythe (1974) rejected the Box-Scheffé test out of hand for this reason.

Fortunately, Brown and Forsythe (1974) reported a robust alternative test for spread that was not subject to the randomization problem. In their procedure, transformed observations are defined as $X_{ijk} = |Y_{ijk} - mdn_{jk}|$, where $i = 1, \ldots n$, and mdn_{jk} is the cell median. Then conventional ANOVA procedures are applied on the X_{ijk}. The Brown and Forsythe technique is more closely related to the average deviation, defined as $AD_{jk} = \Sigma |Y_{ijk} - mdn_{jk}|/n_{jk}$, than to the variance. As such, it should be less influenced by the presence of a few outliers. Unfortunately, derivations of power curves and other properties of the Brown and Forsythe procedure are mathematically intractable.

Another alternative suggested by Brown and Forsythe (1974) is the use of absolute deviations from a trimmed mean, here $X_{ijk}' = |Y_{ijk} - \bar{Y}_{tjk}|$ is used in a conventional ANOVA, where $\bar{Y}_{tjk}$ is the trimmed mean for that cell. This is similar to the Levene (1960) test recommended by Glass (1966), except that Brown and Forsythe use a 10% trimmed mean for which the highest 5% and lowest 5% of the cases are dropped from each sample when computing $\bar{Y}_{tjk}$.

Another alternative is the jackknife test (Miller, 1968), which also subdivides the data into subgroups but has the virtue that all users will obtain the same results, since the subgroups are exhaustive. The subgroups are divided into n subgroups of $n - 1$ observations, that is, one observation is dropped in each subgroup. Then pseudovalues $p_{ijk} = n\, ln\, s_{jk}^2 - (n - 1)\, ln\, s_{ijk}^2$ are defined, where $s_{ijk}^2 =$ the

unbiased sample variance with the ith observation dropped. These pseudovalues are entered into the two-factor ANOVA as the raw data. Prior literature (Brown & Forsythe, 1974; Layard, 1973; Miller, 1968; Martin & Games, Note 1) suggests, however, that on leptokurtic populations, the jackknife test has a Type I error slightly in excess of the nominal alpha. Values of 8%–10% are encountered when $\alpha = .05$.

A virtue of all the present formulations is that they permit multifactor designs, trend tests, or the use of multiple comparisons that are familiar to most ANOVA users. By transforming variance problems into problems of location, they permit a great increase in the type of hypotheses that may be tested (Games, 1978).

Monte Carlo Study

A computer program was written that computed each of the six tests on main and interaction hypotheses on a two-factor independent-groups design. The design was specified as an independent-groups factorial design with four levels of A and three levels of B. The Overall and Woodward (OW) test was formulated using conventional two-factor ANOVA logic, thus taking Z deviations from the observed grand mean rather than assuming $\bar{Z} = 0$, since all the other tests would also be using such conventional ANOVA logic. For each design, 16 samples were drawn for each cell using the pseudorandom number generation "shuffle procedure" of Marsaglia, MacLaren, and Bray (1964). Each such design was replicated 5,000 times, yielding an estimate of FWI or empirical power for $\alpha = .05$ and for $\alpha = .01$. This procedure was repeated using a normally distributed population and also using a population of chi-square values with two degrees of freedom (χ^2_2). The χ^2_2 values were obtained by adding the squares of two independent unit-normal variates.

When the probability of an FWI was assessed, all population variances were set equal to one. For the power comparisons, the variances were specified as in Table 1. In Table 1, the null is false for both main effects, but the null is true for the interaction of the variances. However, due to the use of curvilinear transformations, the interaction null is false for all but the Brown and Forsythe (BF) tests. This is an intrinsic consequence of curvilinear transformations: If additivity is present in the original data, it usually will not be present in the transformed data. Thus the interactions will be excluded when comparing the six tests on power.

Table 1
Values of the Cell Variances (σ_{jk}^2), Row Variances ($\sigma_{.k}^2$), and Column Variances ($\sigma_{j.}^2$) for a Two-Factor Independent-Groups Design

Row factor A_j	Column factor B_k: B_1	B_2	B_3	$\sigma_{.k}^2$
A_1	10	8	6	8
A_2	9	7	5	7
A_3	7	5	3	5
A_4	6	4	2	4
$\sigma_{j.}^2$	8	6	4	

The values obtained for the .05 level of significance and the .01 level of significance showed comparable results, thus only the .05 results are presented in Table 2.

Familywise Risk of Type I Error

The FWI values for the .05 alpha are shown in Table 2. With normally distributed populations, all of the tests show values of FWI that are reasonably close to alpha. The Bartlett and Kendall (BK) and Brown and Forsythe absolute deviation from the trimmed mean (BF_M) yield FWIs slightly larger than alpha, with mean FWI values of .064 and .057, respectively. The jackknife (JK) and Brown and Forsythe absolute deviation from the median (BF_{mdn}) are slightly conservative with mean FWI values of .042 and .036, respectively.

Under the χ^2_2 distribution, the BK and OW tests show the expected extreme FWI values over .05. The BF_M test also shows an inflated FWI approximately three times α, whereas the JK again shows a slight inflation to an FWI $\approx$ .089. Only the Box-Scheffé (BS) and BF_{mdn} tests show excellent control of FWI when the populations are leptokurtic and skewed.

Power

Since all the tests show reasonable control of FWI when the populations are normally

Table 2
Empirical Type I Error Probabilities for a Two-Factor Independent-Groups Design

	Distribution											
	Normal, N (0, 1)						Skewed leptokurtic, χ^2_2					
Sample size	BK	OW	JK	BF_M	BF_{mdn}	BS	BK	OW	JK	BF_M	BF_{mdn}	BS
$n_{jk} = 16$												
A	.066*	.052	.042*	.056	.032*	.047	.499*	.474*	.089*	.145*	.044	.056
B	.062*	.053	.041*	.056	.039*	.047	.403*	.383*	.085*	.122*	.046	.051
AB	.071*	.046	.036*	.059*	.031*	.047	.699*	.664*	.102*	.193*	.048	.055
$n_{jk} = 25$												
A	.055	.049	.044	.057*	.036*	.046	.507*	.498*	.086*	.149*	.048	.053
B	.060*	.053	.044	.054	.038*	.047	.404*	.394*	.077*	.130*	.043	.047
AB	0.69*	.057*	.047	.059*	.039*	.049	.731*	.710*	.095*	.208*	.045	.051
	Mean familywise risk of Type I error											
	.064	.052	.042	.057	.036	.047	.540	.520	.089	.158	.046	.052

Note. BK = Bartlett and Kendall; BS = Box-Scheffé; BF_M = Brown and Forsythe absolute deviation from the trimmed mean; BF_{mdn} = Brown and Forsythe absolute deviation from the median; JK = Jackknife; OW = Overall and Woodward. When $n_{jk} = 16$, the subsample size for the BS test was 4, and the BF_M test used a 12.5% trimmed mean. When $n_{jk} = 25$, the subsample size for the BS test was 5, and the BF_M test used an 8.0% trimmed mean. For all probabilities, $\alpha = .05$.
* $p < .05$ for deviation from the expected familywise risk of Type I error.

distributed, all can be compared for power. As expected, the two tests based on classical theory, the BK and OW, are more powerful than any of the more robust tests. As predicted the BK is always more powerful than the OW. Of the two tests that always provide good control of FWI, the BF_{mdn} is always more powerful than the BS.

Under the χ^2_2 populations, only tests that provide reasonable control of FWI are included. Although the JK is of borderline status in that FWI was slightly inflated, which inflates power also, it was included in Table 3 for comparison to the BS and BF_{mdn}. The power differences between the three tests are relatively small, with the largest difference only .065.

Discussion

If the experimenter is confident that the underlying data are normally distributed, any of the six tests covered could be considered adequate in terms of practical control of FWI. The choice between tests would then be based on power. The tests (under the Normal distribution) are listed from the most powerful to the least powerful (from left to right, respectively) in Tables 2 and 3. There is a gradual reduction as you go from left to right in each row, though many adjacent differences are certainly not significant. Thus, if the experimenter is confident of normality, the BK test is best, although the difference between the BK and the OW test is not large.

However, the experimenter often has little grounds for confidence in the shape of the underlying distributions, particularly with small to medium *n*s. If the populations are leptokurtic, the BK and OW tests can have absurdly large FWI values that far exceed the nominal alpha. Similarly, this study has confirmed Brown and Forsythe's (1974) finding that their BD_M test is not robust when skewed populations are encountered. Only three tests remain relatively robust under all forms of populations studied to date. Of these, the JK is the most powerful, but it is accompanied by a slight rise in FWI under leptokurtosis, sometimes reaching an FWI as large as two alpha. The BF_{mdn} test is next in power but has the disadvantage that it is a test of average deviations rather than variances, and Fellers (Note 2) suggests that it is erratic for *n*s as

Table 3
Empirical Power Probabilities for a Two-Factor Independent-Groups Design

Sample size	Normal, N (0, 1) BK	OW	JK	BF_M	BF_{mdn}	BS	Skewed leptokurtic, χ^2_2 JK	BF_{mdn}	BS
$n_{jk} = 16$									
A	.672	.590	.546	.538	.441	.378	.281	.237	.216
B	.755	.686	.656	.642	.562	.483	.307	.291	.259
$n_{jk} = 25$									
A	.874	.818	.817	.756	.694	.649	.366	.368	.330
B	.938	.898	.910	.852	.817	.756	.429	.465	.418

Note. KB = Bartlett and Kendall; BS = Box-Scheffé; BF_M = Brown and Forsythe absolute deviation from the trimmed mean; BF_{mdn} = Brown and Forsythe absolute deviation from the median; JK = jackknife; OW = Overall and Woodward. When $n_{jk} = 16$, the subsample size for the BS test was 4, and the BF_M test used a 12.5% trimmed mean. When $n_{jk} = 25$, the subsample size for the BS test was 5, and the BF_M test used an 8.0% trimmed mean. For all probabilities, $\alpha = .05$.

small as 5. The BS test deals more directly with variances as such and always has an FWI $\approx \alpha$, but it is lowest in power and has the disadvantage that different random subsamples might yield different outcomes.

Thus there is no one test that can be universally recommended without qualification. O'Brien (1978, in press) gives reasons why the power relations of the several tests vary with the kurtosis of the populations. O'Brien recommends the BF_{mdn} test and the use of the JK on s^2 directly (rather than on $ln\ s^2$ as in the present study). The authors note that platykurtosis is rare in the behavioral data we have seen, since skewed and/or leptokurtic data is more common. If the experimenter is reasonably confident that the data are not platykurtic, the minimal computations needed for the BK test are a reasonable first step.

If this test is not significant, the null hypothesis is retained. However, if the BK is significant, there is the possibility that the result is a Type I error due to leptokurtosis in the population. Thus it is desirable to also reject the null hypothesis by one of the robust tests before making strong interpretations about heterogeneous spread. This two-step process is not ideal, but neither are any of the tests investigated to date.[1]

[1] A FORTRAN computer program for the JK, BF_{mdn}, or BS tests may be obtained by sending a computer tape to the first author.

Reference Notes

1. Martin, C. G., & Games, P. A. *Selection of subsample sizes for the Bartlett and Kendall test of homogeneity of variance.* Paper presented at the meeting of the American Educational Research Association, Washington, D. C., April 1975.
2. Fellers, R. R. *The effects of nonnormality and sample size on the robustness of tests of homogeneity of means.* Paper presented at the meeting of the Northeast Educational Research Association, Liberty, New York, April 1972.

References

Appelbaum, M. I., & Cramer, E. M. Some problems in the nonorthogonal analysis of variance. *Psychological Bulletin*, 1974, *81*, 335–343.

Bartlett, M. S. Properties of sufficient and statistical tests. *Proceedings of the Royal Society of London*, 1937, *160*, 268–282.

Bartlett, M. S. The use of transformations. *Biometrics Bulletin*, 1947, *3*, 39–52.

Bartlett, M. S., & Kendall, D. G. The statistical analysis of variance heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society*, 1946, *8*, 128–138.

Birch, H. G., & Lefford, A. Visual differentiation, intersensory integration, and voluntary motor control. *Monographs of the Society for Research in Child Development*, 1967, *32*(2, Serial No. 110).

Block, J. H. (Ed.). *Mastery learning: Theory and practice.* New York: Holt, Rinehart & Winston, 1973.

Box, G. E. P. Non-normality and tests on variance. *Biometrika*, 1953, *40*, 318–335.

Box, G. E. P., & Anderson, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society*, 1955, *27*, 1–26.

Brown, M. B., & Forsythe, A. B. Robust tests for homogeneity of variance. *Journal of the American Statistical Association*, 1974, *69*, 364–368.

Carlson, J. E., & Timm, N. H. Analysis of nonorthogonal fixed-effects designs. *Psychological Bulletin*, 1974, *81*, 563–570.

Fisher, R. A., & Yates, F. *Statistical tables for biological, agricultural, and medical research.* New York: Hafner, 1963.

Games, P. A. Multiple comparisons of means. *American Educational Research Journal*, 1971, *8*, 531–565.

Games, P. A. Computer programs for robust analyses in multifactor analysis of variance designs. *Educational and Psychological Measurement*, 1975, *35*, 147–152.

Games, P. A. A three-factor model encompassing many possible statistical tests on independent groups. *Psychological Bulletin*, 1978, *85*, 168–182.

Games, P. A., Winkler, H. B., & Probert, D. A. Robust tests for homogeneity of variance. *Educational and Psychological Measurement*, 1972, *32*, 887–909.

Gartside, P. S. A study of methods for comparing several variances. *Journal of the American Statistical Association*, 1972, *67*, 342–346.

Glass, G. V. Testing homogeneity of variance. *American Educational Research Journal*, 1966, *3*, 187–190.

Johnson, E. S., & Baker, R. F. The computer as experimenter: New results. *Behavioral Science*, 1973, *18*, 377–385.

Layard, M. W. J. Robust large-sample competitors for homogeneity of variances. *Journal of the American Statistical Association*, 1973, *68*, 195–198.

Levene, H. Robust tests for the equality of variance. In I. Olkin (Ed.), *Contributions to probability and statistics*. Stanford, Calif.: Stanford University Press, 1960.

Levy, K. J. An empirical comparison of the *Z*-variance and Box-Scheffé test for homogeneity of variance. *Psychometrika*, 1975, *40*, 519–524.

Marascuilo, L. A. Large-sample multiple comparisons. *Psychological Bulletin*, 1966, *65*, 280–290.

Marsaglia, G., MacLaren, M. D., & Bray, T. A. A fast procedure for generating normal random variables. *Communications of the Association of Computing Machines*, 1964, *7*, 4–10.

Miller, R. G., Jr. Jackknifing variances. *Annals of Mathematical Statistics*, 1968, *39*, 567–582.

O'Brien, R. G. Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika*, 1978, *43*, 327–342.

O'Brien, R. G. An improved ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, in press.

Overall, J. E., & Woodward, J. A. A simple test for heterogeneity of variance in complex factorial designs. *Psychometrika*, 1974, *39*, 311–318.

Scheffé, H. A. *The analysis of variance*. New York: Wiley, 1959.

Scheffé, H. A. Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*, 1970, *65*, 1501–1508.

Skinner, B. F. Teaching machines. *Science*, 1958, *128*, 969–977.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Received April 12, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 985–1010

The Role of Fear in Theories of Avoidance Learning, Flooding, and Extinction

Susan Mineka
University of Wisconsin—Madison

The course of fear conditioning and extinction in the avoidance-learning context is complex. This article summarizes the major lines of evidence that demonstrate a dissociation or desynchrony between measures of fear and avoidance responding. The evidence bearing on the role of fear in theories of avoidance learning and extinction is reviewed and critically evaluated. In addition, research is discussed regarding the determinants of fear over the course of avoidance acquisition, flooding, and extinction. Particular emphasis is placed on discussing the extent to which fear extinction is necessary and/or sufficient for avoidance response extinction both with conventional extinction procedures and with response prevention techniques.

In animals and humans fear has long been assumed to play an important role in the mediation of avoidance behaviors that in turn have often been assumed to underlie a variety of neurotic behaviors. The avoidance behavior that frequently accompanies a state of fear has actually been considered by some theorists to be one of the response systems inherent in our definition of fear itself. Lang (1968, 1971), for example, has argued that fear is a complex construct that in humans includes at least three different response systems—verbal/cognitive (subjective), motor (behavioral avoidance), and psychophysiological. These three response systems do not always covary together, and treatments designed to reduce so-called fear may, at least initially, affect one system but not the others. (See also Hodgson & Rachman, 1974; Rachman & Hodgson, 1974.)

Other theorists, whose primary interest has been avoidance behavior in infrahuman organisms, have attempted to understand the role that fear plays in mediating the acquisition, maintenance, and extinction of learned avoidance responses. Over the past 15 years or so, it has become increasingly apparent that the role of fear in mediating any of these facets of learned avoidance behavior is at best not a simple one. In particular, there is often a marked dissociation between fear and avoidance responding (Riccio & Silvestri, 1973) that has led a number of theorists to question whether fear plays any role at all in mediating avoidance responding (e.g., Herrnstein, 1969; Hineline, 1977).

This article has three goals. First, after a discussion of the measurement of fear in animals, there is a brief review of the evidence on dissociation or desynchrony between fear and learned avoidance behavior. Given this background, the second goal is to evaluate comprehensively and critically the role that fear plays in various theories of avoidance acquisition, maintenance, and extinction. Because the majority of the work in this area has centered on the question of the role that fear extinction plays in mediating avoidance response extinction through response prevention or flooding techniques, particular emphasis is placed on work in that area. The third goal of this article is to review, where pertinent evidence is avail-

---

This work was supported by Grant MH-27156 from the National Institute of Mental Health. The author would like to thank Robert Hendersen and Antonio Gino for their very helpful comments on earlier versions of this manuscript.

Requests for reprints should be sent to Susan Mineka, Department of Psychology, University of Wisconsin, 1202 West Johnson Street, Madison, Wisconsin 53706.

able, what determines the course of fear acquisition and extinction in avoidance-learning contexts. In particular, attention is focused on how the avoidance context influences the dynamics of fear conditioning and extinction.

Measurement of Fear

Conditioned fear emerges as the result of pairing a neutral stimulus with an aversive or noxious unconditioned stimulus. Over the past 50 years, numerous response systems have been shown to be sensitive to the effects of aversive conditioning procedures: defecation, heart rate, suppression of ongoing consummatory or operant appetitive behavior, learning a response to escape from a fearful stimulus, passive avoidance of a fearful stimulus, facilitation of ongoing operant avoidance behavior, and so forth. Each one of these indices of conditioned fear has been more or less extensively validated by determining whether the degree of fear as measured by that index is sensitive to conditioning parameters that would be expected on some a priori basis to affect the level of fear that is conditioned (e.g., intensity of the aversive stimulus, number of conditioned stimulus – unconditioned stimulus [CS–US] pairings, etc.). These validation procedures have, of course, each assumed that there is some degree of correlation between the magnitude of the observable response and the internal state of fear itself (McAllister & McAllister, 1971).

Experiments in the area of fear and avoidance have tended to use one of four of the indices of conditioned fear mentioned previously. Probably the most widely used index of the results of an aversive conditioning procedure has been the conditioned emotional response (CER) index, first developed by Estes and Skinner (1941). They observed that hungry rats trained to barpress for food reinforcement decreased their rate of responding when a warning signal was presented for an impending aversive event (usually electric shock). Estes and Skinner labeled this pattern of suppression of ongoing operant appetitive behavior "conditioned anxiety," and over the past 35 years, this phenomenon has been shown in numerous studies to be a reliable and sensitive index of aversive conditioning in a variety of different species (Davis, 1968). The term "conditioned emotional response" was first used by Hunt and Brady (1951) to describe this phenomenon, and it is now the term most widely used to describe this index of conditioned fear or anxiety, two terms that are most often used interchangeably by experimental psychologists. Although others have used the term CER in a more general way to describe the state that has presumably been conditioned in an aversive conditioning procedure, no matter how it is being measured (e.g., heart rate, suppression of operant appetitive behavior, facilitation of avoidance behavior, etc.), in this article the term will be used only to refer to this one particular index of conditioned fear, that is, suppression of operant appetitive behavior. It should be emphasized that in spite of the wide use of the CER as an index of fear or aversive conditioning, there is as yet no consensus as to why positively reinforced behavior is suppressed during a stimulus paired with an aversive stimulus such as shock. (See Blackman, 1977, for one current discussion of this topic.)

A second and closely related index of aversive conditioning used in some of the more recent experiments described later is that developed by Sidman, Herrnstein, and Conrad (1957), Herrnstein and Sidman (1958), and Rescorla and LoLordo (1965). These investigators found that when stimuli paired with shock were presented to monkeys or dogs who were responding on a Sidman (unsignaled) avoidance baseline, the rate of avoidance responding increased dramatically for the duration of the warning stimulus. This facilitation or energization of an operant maintained by negative reinforcement has subsequently been used extensively as an index of conditioned excitatory and inhibitory states based on aversive reinforcers (e.g., Rescorla & Solomon, 1967; Weisman & Litner, 1969, 1972). Scobie (1972) and Morris (1974) have even suggested that it may be a more sensitive index of fear or aversive conditioning than the CER because it can sometimes detect evidence of conditioning when the CER does not.

It should be noted here that the same warning stimulus can be used to produce either suppression or facilitation of ongoing operant behavior, depending on whether that operant is maintained by positive or negative reinforcement. For example, Sidman (1958) reported that monkeys that were maintained on a concurrent chain-pulling response for food reinforcement and lever-press response for shock avoidance showed suppression of the chain-pull response and facilitation of the lever-press response during a warning stimulus for shock. Such results present one of the strongest lines of evidence that there is indeed some central mediating state that is being conditioned in an aversive conditioning procedure. Out of convenience, many experimental psychologists have chosen to use the word *fear* to describe that state, which may manifest itself through a facilitation or suppression of ongoing behavior (Rescorla & Solomon, 1967). In addition, such results underscore the point that it cannot be suppression or facilitation per se that is being conditioned in such experiments; rather, it seems that some central state is being conditioned, the motivating properties of which manifest themselves differently according to which other motivating state is maintaining the operant behavior.

A third index of fear used in some of the experiments reported later is the passive avoidance of a place in which aversive conditioning trials have previously occurred. When given a choice, animals will tend to avoid any such stimulus or place previously paired with an aversive stimulus. To maximize the incentive to approach a fearful place, animals are often food deprived prior to the test, and their latency to enter and eat in the fearful place is measured. The fourth and final index of fear used by some investigators in this area is the conditioned heart rate response. There appears to be some variability between species in the nature of this response, since rats tend to show heart rate decreases, whereas dogs and monkeys tend to show heart rate increases (Brady & Harris, 1977).

Early observations that there is often a lack of concomitance among these different results of an aversive conditioning procedure date back at least as far as Gantt's (1937, 1953) results which indicated that different components of a response conditioned with shock as the US develop at different rates and persist for different periods of time, and the result is a disharmony or cleavage in behavioral, somatic, and psychophysiological response systems. Gantt used the term schizokinesis to describe this phenomenon that he studied most extensively in dogs. More recently, similar observations have been made in several other species. For example, DeToledo and Black (1966) and Brady, Kelly, and Plumlee (1969) have reported that in an aversive conditioning procedure, rats and monkeys show more rapid acquisition of suppression of operant appetitive behavior (CER) than of a heart rate conditioned response. Together, such results indicate that the search for any one uniquely valid index of fear or aversive conditioning in animals will be futile, just as recent evidence from humans has suggested it should be (e.g., Lang, 1968).

This brief review of the measurement of fear in animals certainly indicates that the task of those interested in the relationship between fear and avoidance responding is not an easy one. Given the wide range of measures that may be sensitive to the effects of an aversive conditioning procedure, theorists interested in this relationship should ideally have at hand data that uses multiple-response measurements. Certainly, results that emerge using one index of fear cannot be assumed to apply if a different index of fear is being used. Unfortunately, however, most investigators in this area have used only one index response at a time in their attempts to understand the role of fear in avoidance behavior. Nevertheless, numerous interesting results have emerged over the past 20 years that do help us understand the extent to which fear[1] does or does not play a role in avoidance-learning contexts. More generally, we will see that even when the role

[1] The term *fear* will be used in the present article to refer to any of the four patterns of results of aversive conditioning procedures that have been described previously.

of fear in mediating avoidance responding is in doubt, the question of what happens to fear over the course of avoidance acquisition, maintenance, and extinction is still an interesting and important one. All the conditions necessary for the acquisition and extinction of fear are automatically present in the avoidance-learning context, and yet the CS–US contingency in this aversive conditioning context is unusual, that is, it is a partial reinforcement schedule with nonreinforced CS events whose abbreviated and varying durations are determined by the avoidance latency of the subject. As is demonstrated, the dynamics of fear conditioning and extinction in this unusual context are more complex than those traditionally seen in more straightforward Pavlovian fear conditioning paradigms.

Dissociation Between Fear and Learned Avoidance Behavior

Interest in the association between fear and avoidance responding probably stems from the early appeal of Mowrer's (1947) two-process theory that ascribes a major role to fear as a motivational state in the acquisition of learned avoidance responses. According to this theory, fear, which is conditioned on early trials of a signalized avoidance procedure, serves to motivate the learning of a response that serves to reduce the fear. Reinforcement for avoidance responding comes from termination of the fear-evoking CS. In this theory fear is necessary to motivate the response; once fear has disappeared or extinguished, the avoidance operant should also extinguish. And given a traditional extinction procedure in which shock is no longer presented, fear extinction should be both necessary and sufficient for avoidance response extinction. Other two-process theorists such as Schoenfeld (1950), Sidman (1953), and Dinsmoor (1954) have not invoked the motivational concept of fear in their theories but rather have maintained that stimuli paired with shock early in avoidance become noxious or aversive; consequently, their removal is reinforcing and the avoidance operant is learned. Although these theorists do not use the motivational concept of fear, the aversive or noxious quality of the stimuli whose removal provides reinforcement for the avoidance response should extinguish according to the laws of classical conditioning. Hence the same basic predictions follow from these theories as from Mowrer's theory regarding the necessary and sufficient conditions for avoidance response extinction.

The relationship between fear and avoidance responding is considerably more complex than these early two-process theories predicted. First, there is now considerable evidence that fear of the CS, as indexed by the CER, becomes attenuated over the course of avoidance learning. Kamin, Brimer, and Black (1963) and Starr and Mineka (1977) found that rats trained to a criterion of 27 consecutive avoidance responses (CARs) showed less suppression during the CS than those trained to a criterion of only 9 CARs. Mineka and Gino (Note 1) have further shown that this attenuation of the CER does not occur because the avoidance response itself is about to extinguish. Animals trained to a criterion of 27 CARs show approximately the same resistance to extinction as do animals trained to only 9 CARs. So fear of the CS, as indexed by the CER, is clearly not monotonically associated with the strength of an avoidance response, although it should be noted that no one has yet demonstrated avoidance responding in the complete absence of fear of the CS. Using a different index of fear or stress, Brady (1965) also reported a dissociation between avoidance performance and a physiological correlate of the CER—increases in plasma 17-hydroxycorticosteroid (17-OH-CS) levels. Brady's monkeys showed progressive elevation of 17-OH-CS levels over the first 72 hours of avoidance training, which was then followed by declining levels of 17-OH-CS over the succeeding weeks of avoidance training (Brady & Harris, 1977). Similarly, Coover, Ursin, and Levine (1973) reported that plasma-corticosterone levels in rats were considerably elevated following early avoidance training sessions but that after many (17) training sessions, when avoidance performance was asymptotic, plasma-corticosterone levels only showed small increases over basal levels.

In a somewhat different vein, Rachlin and Herrnstein (1969) have reported a dissociation between the suppressive effects of an aversive shock and its capacity to sustain avoidance. In their second experiment, they found that pigeons who showed little suppression of keypecking with noncontingent shock (CER procedure) did show pronounced negative choice (avoidance) for that component of the schedule. At a minimum, these results suggest that the rate of responding (essentially the CER index of fear) during a stimulus signaling noncontingent shock is not well correlated with the capacity of that stimulus to sustain avoidance. Such results certainly present a dilemma to theorists who maintain that there is any kind of simple relationship between the capacity of a stimulus to sustain avoidance and its capacity to suppress operant appetitive responding.

A second prediction of two-process theory regarding fear and the extinction of avoidance responding has also not been substantiated. In particular, fear extinction does not appear to be necessary for the extinction of the avoidance operant. Although Black (1959) did find that the heart rate conditioned response (CR) extinguished considerably more rapidly than the avoidance response in dogs, he found no significant correlation between the speed of extinction of the cardiac CR and the avoidance response. Using the CER rather than heart rate as an index of fear, Kamin et al. (1963), by contrast, found evidence of avoidance response extinction in the absence of much fear extinction. Rats that were extinguished to a moderate criterion (5 consecutive failures to respond) were still quite fearful of the CS as indexed by the CER, and animals that were extinguished to a more stringent criterion (20 consecutive failures to respond) still showed nonzero levels of fear. Other investigators who used flooding or response prevention techniques to hasten rapid extinction of an avoidance response have reached similar conclusions regarding the lack of necessity of fear extinction in producing avoidance extinction. Coulter, Riccio, and Page (1969), for example, have shown that animals whose avoidance response has been extinguished following flooding are more fearful than animals extinguished with a conventional extinction procedure. In addition, Mineka and Gino (1979a) have shown that an amount of response prevention sufficient to hasten avoidance response extinction does not reduce fear of the CS. So with flooding, as with conventional extinction procedures, fear extinction is not necessary for extinction of the avoidance operant.

This brief summary highlights the evidence showing dissociation between fear and learned avoidance behavior (see also Hodgson & Rachman, 1974; Rachman, 1976; Rachman & Hodgson, 1974; Riccio & Silvestri, 1973). Considerably more discussion and elaboration of this evidence, as well as of the determinants of this dissociation, is made in the remainder of the article. We now turn to a discussion of the role of fear in various theories of avoidance acquisition and maintenance. Attention is focused on how some of these theories have evolved to handle this evidence on dissociation between fear and avoidance.

Fear in Avoidance Acquisition and Maintenance

The Role of Fear in Various Theories of Avoidance Learning

Of all the theories discussed here, Mowrer's (1947) two-process theory clearly ascribes the most important role to fear at various stages of avoidance acquisition, maintenance, and extinction. This theory may, in fact, provide a plausible account for how an avoidance response is initially learned, but it is the phenomenon of the persistence of learned avoidance responses that has most intrigued and plagued learning theorists for the past 25 years. And so, although two-process theory in its various forms has dominated both the theorizing and the experiments done in the field of avoidance learning, its most serious shortcomings have consistently been in its inability to explain satisfactorily the high resistance to extinction of well-learned avoidance responses (e.g., Solomon, Kamin, & Wynne, 1953). The dilemma for two-process theory is that

after dozens or hundreds of consecutive avoidance responses, the source of reinforcement for responding is no longer apparent because each successful avoidance trial constitutes a Pavlovian extinction trial. Hence the fear CR should gradually extinguish, thus removing CS termination as a possible source of reinforcement. After that, the avoidance operant should proceed to extinguish. But as we have already seen, even when the fear CR does become attenuated, the avoidance operant does not immediately begin to extinguish (Kamin et al., 1963; Starr & Mineka, 1977; Mineka & Gino, Note 1).

Solomon and Wynne (1954) attempted to rescue two-process theory from this dilemma by adding to it the two principles of anxiety conservation and the partial irreversibility of the conditioned fear response learned with traumatic shock. These two principles have not proved very useful, however, in explaining the results of experiments demonstrating high resistance to extinction of avoidance responses learned with only moderate levels of shock (Brush, 1957) and conversely the relatively rapid extinction of the CER, even with traumatic shock (Annau & Kamin, 1961). Rescorla and Solomon (1967) further revised two-process theory to account for the apparent lack of fear in the well-trained animal (e.g., Kamin et al., 1963). They postulated that fear is a central state and therefore that lack of concomitance between peripheral measures of fear and avoidance responding does not constitute evidence against two-process theory. This central state of fear, however, is subject to the normal laws of Pavlovian conditioning, including extinction. Thus it is still unclear why this central state of fear does not extinguish in the well-trained animal, and if it does, where the motivation and reinforcement for continued responding come from.

Konorski (1948) and Soltysik (1963) attempted to explain why the fear CR should not extinguish by postulating that it is protected from extinction by the avoidance response that becomes a CS− for shock (inhibitory CS predicting no shock). One problem with this explanation is that it is based on protection-from-extinction experiments done with appetitive responses; no one has demonstrated protection from extinction when the CS− follows the CS+, as in the avoidance case (see LoLordo & Rescorla, 1966; Seligman & Johnston, 1973). Furthermore, the theory explains why fear should persist and provide motivation for responding. However, the evidence discussed previously (e.g., Kamin et al., 1963; Starr & Mineka, 1977) suggests that fear attenuates over the course of avoidance training.

The more recent safety signal or positive reinforcement revision of two-process theory also assumes that the avoidance response becomes a CS− for shock (Bolles, 1970; Weisman & Litner, 1969, 1972). However, this theory does not require that the CS− protect the CS+ from extinction because the CS− assumes the role of a positive reinforcer. The animal continues to make avoidance responses because the response itself (CS−) becomes a positive reinforcer, and fear is not necessary for continued motivation of the avoidance response once the response has become a good CS−. The chief problem with this explanation is that no one has demonstrated that a CS− continues to be a positive reinforcer when it is no longer presented in a fear-eliciting situation (Grossen, 1971; LoLordo, 1969; Seligman & Johnston, 1973). Thus the safety-signal account must posit that the CS+ was protected from extinction to account for the maintenance of the avoidance response as a positive reinforcer (CS−). Until independent evidence exists that residual fear of the CS+ or of the situational cues remains during asymptotic avoidance when the CS− is a positive reinforcer, the safety signal account of the extreme persistence of avoidance responding in extinction is incomplete.

Other theorists (e.g., D'Amato, 1970; Herrnstein, 1969; Hineline, 1977) have further de-emphasized the role of fear in avoidance learning and have instead emphasized the CS's role as a discriminative cue, which "sets the occasion" for responding. The response is presumed to be reinforced by a reduction in shock frequency rather than by reduction of fear. One of these theorists' strongest lines of evidence against the role of the CS as the motivational mediator of an avoidance response comes from the results of

several experiments that attempted to determine if an animal could learn to avoid a CS or discriminative stimulus (S^D) for an avoidance response. For example, Sidman (1955) pretrained cats and rats on an unsignaled shock-delay procedure and then introduced a 5-sec preshock cue that could be delayed or removed. He found that most of the responses occurred during the cue and that the 25%–30% of the responses that did occur in the absence of the cue mostly occurred in postshock bursts. These results suggest that although the removal of a cue can maintain responding, its delay apparently cannot. From the traditional two-process standpoint, one might expect that the animals would learn to avoid a conditioned aversive stimulus if its removal acts as a reinforcer, and yet this does not appear to happen. These results and others summarized by Hineline strongly suggest that the CS in a discriminative avoidance procedure "cannot be seen as simply providing a classically conditioned surrogate for the shock, for several effects are independent of its relation to shock" (p. 396). These theorists do not deny that Pavlovian conditioning of fear may go on in avoidance training and may even affect the rate of responding (Rachlin & Herrnstein, 1969, p. 90), but they do strongly assert that such "classically conditioned responses are not a requirement for the instrumental behavior" (Herrnstein, 1969, p. 61). These theories are, therefore, relatively uninterested in what happens to fear over the course of avoidance learning, although they would probably maintain that the determinants of any fear that does exist should be the result of no more than just the Pavlovian contingencies inherent in the procedure.

Seligman and Johnston (1973) have recently proposed a cognitive theory of avoidance learning that ascribes a role to fear only in the initial stage of training. In the early phase of learning, fear is conditioned to the CS and may be involved in the elicitation of responses. Gradually, however, two expectancies are acquired that serve to maintain the response: an expectancy that a response will be followed by no shock and an expectancy that no response will be followed by shock. If no shock is assumed to be preferred to shock, these two expectancies are sufficient to maintain responding even after fear has extinguished. This theory then predicts that fear and avoidance behavior will not always be well correlated. It further assumes that the degree of residual fear at any phase will be a function of how many Pavlovian fear extinction trials have occurred as a result of successful avoidance responses.

The Determinants of Fear Over the Course of Avoidance Learning

These newer theories of avoidance learning were developed partially as a result of the inability of two-process theory to explain satisfactorily the dissociation or desynchrony between fear and avoidance responding. Each of these theories postulates some new mechanism other than fear to explain the persistence of avoidance responses, for example, the discriminative role of the CS in setting the occasion for a response that results in a reduction of shock frequency or the role of response-outcome expectancies in motivating responses that produce preferable outcomes (i.e., no shock). An explanation of why there is a marked dissociation between fear and avoidance behavior has received only cursory attention (e.g., Seligman & Johnston's, 1973, assumption that fear extinguishes as a result of a simple Pavlovian fear extinction process). However, even if fear does not play a central mediating role in avoidance, the avoidance training procedure automatically programs the necessary and sufficient conditions for the Pavlovian conditioning of fear. Therefore, the question of what happens to that fear over the course of avoidance maintenance and extinction remains one of enormous practical and theoretical importance at least for those interested in fear, even if not for those interested in avoidance.

In a preliminary attempt to study the determinants of fear over the course of avoidance learning, Starr and Mineka (1977) tested the hypothesis that attenuation of fear over the course of avoidance learning results from Pavlovian fear extinction. In a replication and extension of Kamin et al.'s (1963) widely cited study that used the CER as an index of fear, Starr and Mineka compared

fear of the CS in rats trained to 3, 9, or 27 CARs (avoidance-learning [AL] groups) with that of their strictly yoked partners (yoked avoidance-learning [YAL] groups), who received the same pattern of CS and US events but had no avoidance response available. The usual attenuation of fear occurred in the well-trained animals. (The AL-27 group showed less suppression than the less well-trained animals—the AL-3 and AL-9 groups.) However, the yoked group (YAL-27), which had received the same Pavlovian sequence of CSs and USs as the AL-27 group, did not show attenuation of fear. These results suggest that the response contingency per se contributes to the fear attenuation that occurs in well-trained animals.

That a simple Pavlovian fear extinction account does not sufficiently explain attenuation of fear is further supported by other results of Starr and Mineka's (1977) first experiment. Comparisons were made with a third set of groups that were yoked only for the excitatory Pavlovian trials that occurred during the course of avoidance training. These yoked fear conditioning (YFC) groups received the same number and pattern of CS–US pairings as did their AL masters, but they received no nonreinforced CSs. None of the YFC groups differed from each other or from any of the YAL groups. The YFC-27 group did, however, show more fear than the AL-27 group. Because the YFC-27 and YAL-27 groups did not differ in fear, whereas the YFC-27 and AL-27 groups did, simple Pavlovian fear extinction cannot account for the attenuation of fear observed in the AL-27 group. The lack of difference between the YAL-27 and YFC-27 groups is particularly striking in light of the fact that the YAL-27 (and AL-27) group had an average of 63% nonreinforced CS trials, whereas the YFC-27 group had only had reinforced CSs.

To determine which aspect of the response contingency accounts for the decline in fear observed in the AL-27 group, Starr and Mineka (1977) performed a second experiment to assess the role of feedback from avoidance responses in producing this attenuation. One group of animals was trained to a criterion of 36 CARs in a paradigm in which each response was followed by an exteroceptive feedback signal (AL-36-FS). A second group of animals was strictly yoked to the AL group (YAL-36-FS), that is, they did not have an avoidance response available, but they did have a feedback stimulus mimicking the response of the AL group. A third group was also yoked to the AL-36-FS group except that they received no feedback signal (YAL-36-NFS). Both groups (AL-36-FS and YAL-36-FS) that received feedback displayed less fear as indexed by the CER than did the group that received no feedback (YAL-36-NFS), so the response contingency per se is not necessary for the attenuation of fear. The feedback from the avoidance response is sufficient to produce the attenuation.

The mechanism through which feedback produces this attenuation of the CER is not clear. One possibility is that less fear is conditioned to the CS when an FS is present (either in the form of an avoidance response, or an exteroceptive signal). This could occur if an FS were to functionally reduce the intensity of the US by partially inhibiting the fear reaction that would otherwise persist for several seconds following US termination. A second mechanism that might account for the role played by feedback in the attenuation of fear is that fear may extinguish faster when an FS is present. An FS that becomes a powerful conditioned inhibitor of fear may reduce the overall level of fear (e.g., Seligman, 1968). Extinction of fear of a CS presented against such a background might proceed at a faster rate. Although there is as yet no direct evidence bearing on this possibility, it is contrary to the predictions of the Rescorla-Wagner (Rescorla & Wagner, 1972) model that assumes stimuli compete for inhibitory strength. If powerful inhibitory stimuli are already present, this should reduce the amount of inhibitory strength that can accrue to a CS on any extinction trial, so groups with feedback should show less extinction of fear of the CS than groups without feedback. At present, the viability of either of these mechanisms for explaining the role of feedback in producing an attenuation of fear remains to be determined.

Other investigators have also observed attenuation of fear over the course of avoidance learning (e.g., Linden, 1969; Weisman & Litner, 1972), but this is not a universal result. Morris (1974) observed no attenuation of fear in rats trained to 27 CARs when he used a transfer of control test on a Sidman avoidance baseline (Rescorla & LoLordo, 1965); that is, the 27 group showed facilitation of avoidance during the CS equal to that of the 9 group. Morris used an FS following avoidance responses, and this may account for his failure to observe attenuation because attenuation may already have occurred in his AL-9 groups that received an exteroceptive FS, or as discussed earlier, less fear may be conditioned in the first place when an FS is present. Alternatively, different results may be obtained in animals tested for fear with the CER as opposed to the Sidman procedure. The weight of evidence indicates that fear does attenuate with extended avoidance training, although which mechanism produces this attenuation is not yet clear. It should again be noted, however, that fear has never been demonstrated to extinguish completely (e.g., by a CER test or with the Sidman procedure) even with extensive training.

That fear diminishes over the course of a run of consecutive avoidance responses would not be particularly surprising even to a two-process theorist, if the avoidance response concurrently became weaker. If this were the case, one would not be so inclined to speak of a dissociation between fear and avoidance. Recent results of Mineka and Gino (Note 1), however, indicate that the avoidance response is not weaker after a run of 27 consecutive responses than after a run of only 9 responses. Animals trained to a criterion of 27 CARs, in a situation comparable to that of Starr and Mineka (1977), are equally resistant to extinction as animals trained to a criterion of 9 CARs. As indicated by Mackintosh (1974, p. 334), this result creates a serious theoretical problem for theorists who argue that fear motivates the responding of a well-trained animal. By contrast, these results do not present a problem for other theorists who postulate other sources of motivation or reinforcement for avoidance.

Fear and the Extinction of Avoidance Responses

The Role of Fear Extinction in Traditional Avoidance Extinction

The dissociation between fear and avoidance behavior discussed so far has been the diminution of fear as the avoidance response becomes better learned. A two-process theorist might assume that fear would ultimately diminish to a sufficient extent to cause extinction of the avoidance response. As demonstrated earlier, the traditional two-process theory of avoidance in fact assumes that fear extinction precedes and determines avoidance response extinction, but this simple analysis is not correct: The direction of the dissociation between fear and avoidance behavior is not always the same. Rather, as extinction of the avoidance response begins, fear of the CS remains fairly intense. Kamin et al. (1963) reported that animals at a moderate extinction criterion (five consecutive failures to respond) were nearly as fearful of the CS as indexed by the CER as animals that had received no extinction trials. Clearly, animals who have reached a moderate avoidance extinction criterion have not done so because their fear has extinguished. Animals must begin to stop responding for some reason other than that their fear has extinguished. As extinction of the response proceeds, fear extinction is likely to follow. Unfortunately no single experiment has compared fear attenuation (and extinction) across the course of the acquisition of a well-trained avoidance response and its extinction. It is possible that fear may temporarily increase when the animal begins to cease responding. This possibility is suggested by a comparison of the results of Kamin et al.'s two experiments, which unfortunately are not strictly comparable because of differential delay of the time when the CER test was made. Kamin et al.'s animals who had made five consecutive nonresponses (failures to respond in the CS–US interval) were more afraid of the CS than

were the animals who had made 27 consecutive responses.

This dissociation between fear and avoidance responding during extinction is nowhere more apparent than in the literature on the flooding of avoidance responses. Flooding or response prevention techniques are effective in producing rapid extinction of well-learned avoidance responses, even though such responses are resistant to extinction with conventional extinction procedures. These techniques involve prolonged exposure to the CS, either with the response forcibly prevented or with CS termination noncontingent on the response. A variety of theories has emerged to explain the efficacy of these techniques. As is demonstrated, these theories vary in their ability to accommodate the evidence on dissociation between fear and avoidance responding. No one of these theories can adequately account for all of the relevant data.

The Role of Fear Extinction in Theories of Flooding

Two-Process Fear Extinction Theory

Two-process theory in its various forms predicts that response prevention or flooding techniques should be effective in hastening extinction of avoidance responses. By the two-process account, flooding techniques that allow extended nonreinforced exposure to the CS should assure extinction of the fear CR, thus eliminating both the motivation and the reinforcement for continued responding.

The evidence relevant to this traditional two-process account is mixed. The discussion of this evidence is organized around three general, sometimes overlapping questions. First, to what extent is fear extinction necessary and/or sufficient for a flooding effect? Second, which results of response prevention experiments are difficult to accommodate within a two-process framework? And third, what evidence exists that some process other than fear extinction must also be operating during flooding?

Is fear extinction necessary and/or sufficient? Perhaps the most obvious attempt to demonstrate that fear extinction during flooding is not necessary for avoidance extinction is that of Marrazo, Riccio, and Riley (1974). They argue that fear extinction cannot account for flooding results because a group in their experiment that received reinforced CS presentations during flooding extinguished as rapidly as a group that received nonreinforced CS presentations. Fear could not have extinguished in the former group, so the effect of flooding was not solely attributable to Pavlovian fear extinction. There are, however, two problems with this interpretation. First, Bersh and Miller (1975) showed that Marrazo et al.'s results were due to their use of long (5-sec) shocks during flooding. These long shocks seemed to result in jumping and rearing behavior being conditioned to the situation, which in turn facilitated extinction by serving as incompatible responses. It is important to emphasize, however, that this is not the same process as that involved in regular flooding. When they gave short ($\frac{1}{2}$-sec) shocks during flooding, the avoidance responses were much slower to extinguish than with regularly flooded animals. Second, there are some logical problems with Marrazo et al.'s conclusions. That their two groups (fear conditioning and fear extinction) showed equally rapid extinction of the avoidance response does not imply that they did so for the same reason. Their results show at most that fear extinction is not necessary for avoidance extinction, but it may well be sufficient and may even be a necessary part of the traditional flooding process.

Presenting further problems to a two-process account, a number of investigators have noted that fear of the CS remains following extinction after flooding. Page (1955) and Coulter et al. (1969) showed that rats that reach an extinction criterion following response prevention in a one-way shuttlebox show greater fear of the safe side, as indexed by a passive avoidance test, than do rats reaching the same extinction criterion following a regular extinction procedure. Linton, Riccio, Rohrbaugh, and Page (1970) later demonstrated that a group receiving the response blocking procedure shows a small decrement in fear as compared to nonblocked,

nonextinguished controls, but rats that had received blocking and extinction trials showed even less fear, and rats that had been extinguished in the regular fashion showed the least fear.

There are, however, several reasons why these results of Coulter et al. (1969) and Linton et al. (1970) alone may not destroy the fear extinction account of flooding. First, Mackintosh (1974) has suggested that perhaps only a certain threshold amount of fear must be elicited to motivate an active avoidance response. Flooding may reduce the level of fear below this critical threshold but not enough to abolish fear, as indexed on a passive avoidance test. A recent experiment partially consistent with this idea has been reported by Monti and Smith (1976). They found that flooded rats demonstrated less fear of the CS, as indexed by a CER test, than did control rats who had spent a comparable amount of time in the home cage. The flooded rats did not, however, show zero levels of fear, and the difference between the two groups was significant for only the first three trials of the CER test (see also Corriveau & Smith, 1978). More definitive support for Mackintosh's idea would be provided by a demonstration that this amount of flooding was sufficient to rapidly extinguish the rats' avoidance response as compared to the effect of the home cage control treatment on the avoidance response. Such a demonstration is necessary because recent results of Mineka and Gino (1979a) indicate that an amount of flooding sufficient to reliably hasten the extinction of a well-learned shuttlebox avoidance response (20 trials) is not sufficient to reduce fear of the CS. A greater amount of flooding (30 vs. 20 trials) does reliably hasten avoidance extinction and reduce fear of the CS (Mineka & Gino, 1979a). At present, then, there is no good evidence to support Mackintosh's threshold idea: Fear extinction may occur during flooding, but it does not appear to be necessary for avoidance response extinction.

There is, however, a second reason why the Coulter et al. (1969) and Linton et al. (1970) results alone do not destroy the fear extinction account of flooding. Shipley, Mock, and Levis (1971) have criticized the Coulter et al. and Linton et al. experiments because total nonreinforced exposure to the CS was vastly different in blocked, blocked-then-extinguished, and normally extinguished groups. Their results suggest that when amount of nonreinforced CS exposure is held constant, residual fear, as indexed by a passive avoidance test, does not differ across blocked and nonblocked regularly extinguished groups. Furthermore, they found that total CS exposure time was similar in blocked and nonblocked groups when the extinction criterion was finally met. They interpret these results as supporting the two-process explanation that response prevention produces its effect through fear extinction. Baum (1971) also showed that flooded and nonflooded groups that had reached the same extinction criterion recovered equally from extinction when a loud buzzer was presented. He took this to indicate that residual fear in all groups following extinction was roughly equivalent, thereby contradicting Coulter et al.'s conclusion. There are as yet no experiments using CER as an index of fear that compare residual fear in blocked-then-extinguished and regularly extinguished groups. Such experiments are necessary before any definitive conclusions can be reached regarding amounts of residual fear following flooding versus conventional extinction procedures. Shipley et al.'s use of a passive avoidance test as an index of fear confounds behavioral passive avoidance with fear in a situation in which the two groups have had differential opportunity for such behavior to have been reinforced, and Baum's recovery procedure is hardly a validated index of fear. (See also Corriveau & Smith, 1978, for a more complete discussion of these issues.)

Berman and Katzev (1972) have also shown the importance of nonreinforced CS exposure in an experiment that equated CS exposure across five groups. Four groups received one of four different flooding treatments, each following two sessions of two-way shuttlebox avoidance acquisition. The fifth CS-time-control group received as many trials of response-contingent CS termination as were necessary to equate total CS exposure during this treatment phase to that

in the four flooded groups. This latter CS-time-control group extinguished faster than a nontreated control group and did not differ significantly from two of the four flooded groups. (The two groups whose responses were blocked during treatment extinguished faster than the two groups for whom responses were allowed but for whom CS termination was not response contingent.) Berman and Katzev point out that these results suggest that caution is necessary in interpreting the results of nearly all response prevention experiments in which total nonreinforced CS exposure is confounded with the response prevention procedure itself.

The Monti and Smith (1976), Shipley et al. (1971), and Berman and Katzev (1972) results do counter some of the arguments against the fear extinction account of flooding made by Page (1955), Coulter et al. (1969), and Linton et al. (1970) who all found differences in residual fear in flooded and regularly extinguished animals. These results all suggest that total nonreinforced CS exposure—which presumably allows for fear extinction to occur—may be the crucial variable in producing rapid extinction of avoidance. It must be emphasized, however, that this conclusion is based on the as yet unsupported assumption that the amount of fear extinction of an avoidance CS is directly related to the amount of nonreinforced CS exposure. Shipley (1974) did report that this is the case for straightforward fear conditioning and extinction. However, as yet no one has studied this issue directly in the avoidance/flooding situation, that is, whether fear extinction of an avoidance CS (as measured by a CER test) is a simple function of total amount of nonreinforced CS exposure. The results of Starr and Mineka (1977) indicate that the dynamics of fear conditioning in the avoidance situation are not identical to those in a more straightforward fear conditioning situation (cf. Starr & Mineka's yoked groups). In addition, Monti and Smith found that response prevention was more effective in eliminating fear conditioned in a classical paradigm than in eliminating fear conditioned over the course of avoidance learning. And perhaps most important are the results of Mineka and Gino (1979a) which indicate that with a large amount of CS exposure (600 as opposed to 400 sec), the forced aspect of the flooding procedure is more effective in reducing fear of the CS than is an equal amount of the self-exposure that occurs in traditional extinction. So caution is clearly necessary in extrapolating from results which indicate that nonreinforced CS exposure is directly related to avoidance response extinction (Berman & Katzev, 1972; Shipley et al., 1971) to the conclusion that fear extinction is mediating that extinction of the avoidance response. (See also the paradox discussed later presented by the Berman & Katzev, 1972, and Shipley et al., 1971, results.) Such caution seems particularly important given the results of Mineka and Gino (1979a) which indicate that fear as indexed by CER suppression does not diminish, given an amount of flooding that reliably hastens extinction of the avoidance response. Overall, the weight of the evidence seems to indicate that fear extinction is not a necessary part of flooding for avoidance response extinction, although given the results discussed so far, it may well be sufficient.

Other problems for a two-process fear extinction account. A number of other experiments have been considered by some reviewers such as Baum (1970b) to raise problems for the two-process account of flooding. For example, Benline and Simmel (1967) have results which suggest that the effects of flooding procedures may produce only temporary decrements in avoidance responding, perhaps through the learning of competing responses rather than through the extinction of fear. In their experiment, rats that had received 40, 80, or 160 blocking trials over a 5-day period showed response decrements on the first few sessions of extinction as compared to a nonblocked control group. By the 4th and 5th day of extinction, however, the blocked groups were responding as fast and as often as the nontreated control group. These results are difficult to interpret, though, because the control group itself showed no signs of extinction over 5 days (a questionable result considering that there had been only 50 acquisition trials on the 1st day). This makes the meaning of any spontaneous recovery in the blocked groups unclear.

Other experiments (e.g., Polin, 1959; Shearman, 1970) have also given multiple extinction sessions, and they did not see this pattern of spontaneous recovery in flooded groups. Actually, there is no a priori reason why spontaneous recovery of the avoidance response, even if it were convincingly demonstrated, should provide strong evidence against a two-process account of flooding. Spontaneous recovery of any conditioned response can be expected to occur following extinction (Kimble, 1961; Pavlov, 1927). If the conditioned fear response shows spontaneous recovery, then the avoidance response might be expected to recur also.

Potentially more damaging to a two-process account of flooding are the results reported by Werboff, Duane, and Cohen (1964). These investigators found a dissociation between autonomic (heart rate) indices of fear and avoidance responding; rats that had undergone a treatment similar to flooding showed greatly elevated heart rates, even though they were no longer responding. Although Rescorla and Solomon's (1967) version of two-process theory attempts to get around the problem of dissociation between peripheral and other indices of fear by postulating that fear is a central state, their theory has not been taken to postulate that peripheral indices can exist in the absence of the central state (although the reverse can be true) as the Werboff et al. data indicate. At a minimum these data extend the observations of dissociation between fear and avoidance following flooding to include such peripheral (psychophysiological) indices of fear.

Two other lines of research present problems to a two-process account of flooding because the results cannot easily be explained by a two-process account. Lederhendler and Baum (1970) reported that mechanical disruption of their rats' behavior during flooding (when abortive avoidance responses and freezing were occurring) greatly enhanced the efficacy of the flooding treatment as compared to the efficacy of the flooding for normally flooded animals. They interpret their results as supporting relaxation theory discussed later and point out that a two-process account has difficulty explaining the results. Furthermore, Baum (1970a) found that a loud buzzer during flooding enhances the efficacy of the flooding treatment. Since a loud buzzer should, if anything, increase the ambient level of fear, a two-process account of flooding has difficulty explaining why a loud buzzer enhances efficacy of the treatment. It may be that an increase in the ambient level of fear produced by a loud buzzer enhances the efficacy of flooding by habituating the animal to the state of fear itself rather than by extinction of the fear CR. (See Watson & Marks, 1971, for a similar argument for humans.) Baum (1970a) himself argues for a distraction interpretation of the results, which is orthogonal to a two-process account.

Evidence for another process other than fear extinction. Overall, the above experiments indicate that a two-process account of flooding does not account convincingly for all of the relevant data and that fear extinction is not a necessary part of flooding. But even if total nonreinforced CS exposure were the critical variable, as Shipley et al. (1971) and Berman and Katzev (1972) have suggested, a paradox more damaging to the two-process account than the experiments discussed above remains to be resolved: If flooded animals have had more CS exposure and therefore more extinction of the fear CR, then they should (and do) stop responding sooner than nonflooded animals. But if, as Berman and Katzev suggest, extinction of the avoidance response is solely a function of extinction of the fear CR, which in turn is solely a function of total nonreinforced CS exposure, then groups that have met the same extinction criterion should have done so because their fear CR has extinguished equally. This should occur regardless of whether the response has extinguished following flooding or following conventional extinction procedures. Yet the Page (1955), Coulter et al. (1969), and Linton et al. (1970) studies have indicated that groups reaching the same extinction criterion do not demonstrate equal amounts of the fear CR. Their flooded groups may indeed have had less CS exposure and therefore more residual fear than their nonflooded group (cf. Shipley et al., 1971), but their avoid-

ance response did extinguish. Again we see that although complete fear extinction may be sufficient for avoidance response extinction, it does not appear to be necessary. Some other form of learning in addition to extinction of the fear CR must generally also occur during flooding and contribute to rapid extinction of the avoidance response.

Even Berman and Katzev (1972) have some evidence indicating that Pavlovian fear extinction is probably not all that occurs during flooding. Their blocked-spaced trial group did extinguish significantly faster than their CS time control group, which indicates that some other kind of learning occurred in the former group. So if Pavlovian extinction of the fear CR cannot explain all of the relevant data on flooding, what other kind of learning can be taking place during flooding that could account for its efficacy in hastening avoidance extinction?

The safety signal revision of two-process theory must argue that the positive reinforcing properties of the response as a CS− extinguish during flooding. The avoidance response cannot be made during flooding, so reinforcement no longer occurs in the avoidance apparatus. This theory predicts that a CS− established during avoidance learning would no longer serve as a positive reinforcer if the effects were measured following flooding (as contrasted to the results of Weisman and Litner, 1969, when the effects of the CS− were measured before any flooding or extinction trials). No such evidence exists, so the adequacy of this account of flooding cannot be assessed. Some theorists (e.g., Seligman & Johnston, 1973) have even argued that safety signal theory cannot predict that flooding should work because if the response does not occur, how could the CS− properties of its feedback ever extinguish? Furthermore, the complex changes in fear that occur during flooding are of little interest to safety signal theorists because fear is not involved in asymptotic avoidance according to this theory.

Competing-Response Theory

Page (1955), Coulter et al. (1969), and Linton et al. (1970) have argued that what is learned during flooding is a response competing with that learned during acquisition: "Since CS onset results in a fear response which is reduced by fear offset, then any response *S* makes when the CS is terminated will be adventitiously reinforced" (Coulter et al., 1969, p. 380). Although these investigators stay within a two-process framework, the emphasis has changed from what happens to the Pavlovian fear CR to what happens to the instrumental response made to reduce that fear CR. This theory nicely explains the dissociation between the fear CR and avoidance responding. Furthermore, Shipley et al. (1971) have noted that blocked groups show lower activity levels than do nonblocked groups. Shearman (1970) has also noted that all of the 15 of his 40 flooded animals who made no responses over 9 days of extinction also made no intertrial interval responses, which thus indicates low activity levels. These low activity levels could indicate that freezing has become the "competing response." The Marrazo et al. (1974) and Bersh and Miller (1975) experiments discussed previously certainly indicate that competing responses learned during flooding can mediate the extinction of avoidance responding. No experiment has conclusively demonstrated, however, that competing responses either do mediate extinction or are necessary for such extinction. It should be noted that this conclusion is in the same vein as that reached regarding the necessity of Pavlovian fear extinction. As demonstrated earlier, even those experiments (e.g., Bersh & Paynter, 1972) purporting to demonstrate that fear extinction must contribute to avoidance extinction remain inconclusive because the Page (1955), Coulter et al. (1969), Linton et al. (1970), and Mineka and Gino (1979a) data still stand: Without much fear extinction, the avoidance response can extinguish. Analogously here, Black's (1958, 1959) results from flooding done under curare seem to show that a learned competing response is not necessary for extinction to occur: Dogs given no opportunity to learn a competing motor response because they were paralyzed by curare when flooding

was carried out still showed rapid avoidance response extinction.

Relaxation Theory

Baum (1970b) has criticized the competing-response theory on several different counts. His own experiments indicate that "undifferentiated exploratory behavior and grooming" (1970b, p. 281) tend to replace the extinguished avoidance response rather than any specific response such as crouching or freezing. In addition, he notes that a competing-response hypothesis has difficulty explaining why higher shock intensity and overtraining decrease the efficacy of a fixed amount of flooding. One could argue that a better learned avoidance response has more difficulty being overcome by a competing response, but one could also predict that adventitious reinforcement for a competing response in groups trained with higher shock intensity should be greater, thus increasing the efficacy of flooding. Baum's own analysis of what happens during flooding is that the animal learns to relax in the presence of the CS, an idea stemming from Denny's (1971) relaxation theory of avoidance learning. Evidence consistent with Baum's relaxation theory has been provided recently by Hawk and Riccio (1977). These investigators reasoned that if relaxation responses are responsible for avoidance extinction, then a technique that hastens the emergence of relaxation responses should enhance the efficacy of flooding. They presented an independently established CS− (presumably an elicitor of relaxation responses) during flooding for one group, and they did find more rapid extinction in that group. Unfortunately for Baum's theory, a group that received a novel CS during flooding showed rates of extinction comparable to those of the CS− group.

Actually, this relaxation account is similar to the fear extinction account of flooding, except that Baum (1970b) requires that the animal's normal, nonfearful behavioral repertoire (i.e., relaxation responses) return before one assumes that fear has extinguished. Hence the problems with this account are the same as those discussed previously regarding the fear extinction account. Baum (1970b) himself admits that his relaxation theory "still fails to explain evidence of fear (and no relaxation) following response prevention, even though the avoidance response has been extinguished" (p. 282). Furthermore, Morokoff and Timberlake (1971) have results which indicate that relaxation responses are at least not necessary for rapid extinction to occur. Their group that showed the most rapid avoidance response extinction showed six times as many fear responses (evidence of nonrelaxation) as did a group that extinguished substantially more slowly.

Baum concluded his review (1970b) of the three main accounts of flooding by stating that no one of these three accounts seems to explain all the relevant experiments and that all three accounts may be partially correct. His conclusion still stands. In addition, Baum points out that the process most involved in a given experiment (Pavlovian fear extinction, competing-response learning, or active relaxation) may depend on what particular parameters and procedures are used. More direct support of this point is given later.

Cognitive Theory

Another account of how flooding produces its effect has recently emerged as part of Seligman and Johnston's (1973) comprehensive cognitive theory of avoidance learning based on Irwin's (1971) cognitive theory of motivation. As discussed earlier, in this theory conditioned fear plays a role only in the acquisition of the avoidance response; in the well-maintained response, fear no longer plays a motivating role but is replaced by expectancies that responding produces no shock and that not responding is followed by shock and by a preference for no shock as compared to shock. The extreme persistence of avoidance responding in extinction is easily explained by this theory because fear may extinguish without a change in the expectancies that responding is necessary to avoid shock. Typically, the animal does not wait around long enough in the presence of the CS to have its expectancies disconfirmed. According to this cognitive theory of avoid-

ance, flooding hastens extinction of the avoidance response because the animal's expectancy that not responding will lead to shock is disconfirmed, and the more disconfirmation the animal receives, the faster it should stop responding. Seligman and Johnston claim that this theory can easily account for the previously discussed dissociation between extinction of fear and extinction of the avoidance response because disconfirmation of the expectancies governing responding could occur faster than Pavlovian fear extinction. Thus, as in the competing-response account, the emphasis here is that the response may change (i.e., extinguish) in the absence of any change in the level of fear.

At first glance, the attractiveness of this cognitive account of flooding lies in its ability to explain virtually any outcome of a flooding experiment; because extinction of conditioned fear is neither necessary nor sufficient for extinction of the avoidance response, the theory can predict either more or less fear in blocked as compared to normally extinguished groups. The fact that results indicate, if anything, more fear in blocked than in nonblocked groups presents no problem to the account, but neither would the opposite results. Failure of flooding to produce rapid extinction can be attributed to insufficient disconfirmation of the governing expectancies. However, here also lies the problem with this cognitive account of flooding. Because there is no independent way of measuring the governing expectancies, the cost of this flexibility is that it makes this theory impossible to test. This is orthogonal to any criticism of the cognitive theory of avoidance learning as a whole because most existing data on avoidance seem to be interpretable within the theory. (But see earlier discussion of Starr & Mineka, 1977, for one exception.) It certainly does detract, however, from its usefulness as an alternative to other current accounts of how flooding works.

Although formally the cognitive theory of avoidance learning is different from the discrimination theory of D'Amato (1970) and Herrnstein (1969), the latter theory's account of flooding would probably be similar; flooding would probably be seen as working by facilitating a detection of the change in reinforcement contingencies between acquisition and extinction. Additionally, the response might be thought to extinguish because it no longer brings reinforcement in the presence of the S^D. No predictions about fear or fear extinction during flooding would be made. The problems with this account are similar to those with the cognitive account: There is no independent way of assessing whether an amount of flooding that was insufficient to reduce resistance to extinction had failed because the response still produced reinforcement in the presence of the S^D.

Summary

Four accounts of how flooding hastens extinction of avoidance responses have been discussed. None is completely adequate. It is important to recognize that different learning processes may underlie the flooding effects with different kinds of avoidance responses. In particular, fear extinction may be more central to the extinction of some avoidance responses than to others. The myriad of experiments investigating flooding have involved not only different species (dogs for Black, 1958; Carlson & Black, 1959; Solomon et al., 1953; rats for all other studies) and different responses (one-way shuttlebox, jump-up box, two-way shuttlebox) but also responses trained to vastly different criteria (varying from three consecutive avoidance responses with less than 10 training trials to several hundred training trials over many days). In general, experiments taken to support two-process fear extinction accounts have used fairly well-learned two-way shuttlebox responses (e.g., Berman & Katzev, 1972; Monti & Smith, 1976; Polin, 1959; Shearman, 1970; but Mineka & Gino, 1979a, is an exception). Experiments taken to support competing-response accounts have generally used less well-learned one-way shuttlebox responses (e.g., Coulter et al., 1969; Linton et al., 1970; Page, 1955). Baum (1970b) has used moderately well-learned jump-up box responses to support his relaxation theory. Thus direct comparisons among these ex-

periments must be made with extreme caution. Baum has suggested that which account of flooding is most applicable may depend on the particular avoidance response being studied as well as the particular parameters of acquisition, shock intensity, and so forth. With special regard to understanding the dissociation between fear and avoidance, it is important to determine the extent of such dissociation following flooding of different avoidance responses.

The hypothesis that different learning processes mediate flooding effects in different avoidance situations may seem unnecessarily complicated. However, there are other indications of the different natures of two-way shuttlebox learning and one-way or jump-up box learning (e.g., Bolles, 1970; Seligman, Maier, & Solomon, 1971; Stampfl & Levis, 1973; Turner & Solomon, 1962). Bolles (1970), for example, has shown that the CS termination contingency (the source of fear reduction in traditional two-process theories and the source of informational feedback in more recent theories) is not important in one-way avoidance responding, although it is important in two-way shuttlebox responding. Levis and Stampfl (1972) and Stampfl and Levis (1973) have shown that responding to a serial CS is different in the one-way and the two-way situations. In a one-way apparatus, rats respond primarily to the first segment of a CS, although fear of the final segment is higher (Boyd & Levis, 1976). In the two-way shuttlebox, by contrast, rats respond primarily to the second segment of a serial CS. Seligman et al. (1971), in accounting for the failure to demonstrate learned helplessness in the one-way situation, point to the fact that one-way learning can be place learning, whereas two-way shuttlebox learning must of necessity be response learning. With these facts in mind, the idea that different learning processes may mediate the flooding effect in one-way and two-way situations seems reasonable. In two-way response-learning situations, any learning process that affects the motivation or reinforcement for responding (e.g., Pavlovian fear extinction and removal of the response/CS termination contingency) may well be sufficient to hasten extinction of the response. Whether fear extinction is necessary is much less clear given the results of Mineka and Gino (1979a). In the one-way situation, by contrast, reducing the motivation for responding or the response/CS termination contingency during flooding may not even be sufficient to hasten extinction of the response. If, as Seligman et al. (1971) and Bolles (1970) suggest, "running [or jumping] from a dangerous place to a safe place [is] an *innate* response" (Seligman et al., 1971, p. 370), then the animal may need either to learn a competing response or to actively relax in the formerly dangerous place or to have a cognitive change in its expectancy as to what is a safe place. So for one-way responses, fear extinction may be neither necessary nor sufficient for avoidance response extinction.

Nonspecificity of Flooding Effects and Implications for Theories of Flooding

Although there are major differences among the four major theories of flooding, they do have certain elements in common. The Pavlovian fear extinction account and the learned competing-response account are both embedded in a two-process framework, although they differ on which of the two processes is affected during flooding. The relaxation account is similar to the Pavlovian fear extinction account, except that it requires that the animal's normal nonfearful behavioral repertoire have returned before fear is acknowledged to have extinguished. The notion of learning to relax, reminiscent of Wolpe's (1958) reciprocal inhibition theory in which relaxation is learned as a response to previously fear-evoking stimuli also has features in common with a learned competing response, albeit of a different sort than that specified by Page (1955) and Coulter et al. (1969). In addition, the cognitive account has similarities to the competing-response account in that although fear may remain, the response made in the presence of that fear changes, either because of a change in act-outcome expectancies or because of adventitious reinforcement for the new response.

In addition to the above common points, all

four accounts share an implicit assumption that the effects of flooding should be quite specific (Bersh & Keltz, 1971). This specificity should manifest itself in two ways. First, because all four theories are associative in nature, they are silent on the issue of how nonassociative changes could be involved in flooding effects. Second, they all predict that the effects of a flooding treatment should be relatively specific to the avoidance response that has been flooded, that is, they do not predict that a flooding treatment applied to one avoidance response could have transsituational effects, such as hastening the extinction of a second avoidance response learned to a different CS. Both of these assumptions have been challenged recently.

Problems for the first aspect of the specificity assumption are best illustrated by a recent set of experiments reported by Crawford (1977). Crawford found that confinement in novel or fearful places produced nearly as large a facilitation of jump-up response extinction as did a response prevention procedure. She argues that these results are best explained by a species-specific defense response (SSDR) account of flooding (Bersh & Keltz, 1971). This account bears some resemblance to the competing-response account, except it emphasizes that the new response emerges spontaneously when the dominant SSDR is punished or suppressed, not as a result of adventitious reinforcement, that is, it is essentially a nonassociative account of flooding (Bolles, 1972). By this account, during response prevention as well as during confinement in a novel or fearful place, freezing becomes the dominant SSDR, and so when extinction of the avoidance response begins, freezing replaces jumping or fleeing as the dominant response.

Two other lines of evidence also suggest that nonassociative changes, such as a change in the SSDR hierarchy, produce effects that look much like traditional flooding effects. This questions the extent to which flooding acts via associative changes. Baum and LeClerc (1974) found that an irrelevant stress procedure—requiring rats to swim for 5 minutes—produced as large an effect on avoidance extinction as did a regular response prevention procedure. In addition, Monti and Smith (1976) reported that confinement in a shuttlebox alone, with no CS presentations, produced as large an effect on fear (CER) extinction as did a traditional flooding procedure. Together these results suggest that nonassociative changes—perhaps a change in SSDR hierarchy—can produce effects that appear identical to flooding effects.

There is one serious problem, however, in extending this conclusion to the degree that Crawford (1977) has done: Showing that a change in SSDR hierarchy can produce floodinglike effects does not mean that flooding normally acts via this process. For example, Baum and LeClerc (1974) went on to show that their irrelevant stress procedure probably did not produce its effect through the same process as did the response prevention procedure. When a 2-hour time delay was interpolated between the treatment procedures and extinction, the irrelevant stress procedure no longer had an effect, although the response prevention procedure did. This would be expected if irrelevant stress produces its effect through temporary nonassociative changes. Neither Crawford nor Monti and Smith (1976) ran such time delay groups to determine whether their nonspecific procedures were as effective after a delay as were their traditional flooding procedures. So, although there is some indication that nonassociative changes can produce floodinglike effects, it is not yet clear to what extent similar nonassociative changes mediate the effects of conventional flooding procedures.

The second aspect of the specificity assumption mentioned above—that flooding one avoidance response should not hasten the extinction of a second response learned to a different CS—has also been challenged by a recent set of experiments reported by Mineka (1976). In her first experiment, Mineka examined the comparative effectiveness of relevant and irrelevant flooding in hastening the extinction of a two-way shuttlebox response. All rats were trained to perform two different avoidance responses—a one-way jump-up response with a tone CS and a two-way shuttlebox response with a light CS. Flooding the jump-up response hastened extinction of the

shuttlebox response to nearly as large an extent as did flooding the shuttlebox response itself. These results demonstrate that flooding one avoidance response can hasten the extinction of a different response learned to a different CS. In other words, the effects of a flooding experience can be more general than previously thought.

Mineka (1976) then asked whether generalization of fear extinction across CS modalities (Pavlov, 1927) could have mediated this irrelevant flooding effect. By this account, extinction of fear to the tone CS and/or jump-up box grid that occurs during jump-up box flooding generalizes to produce extinction of fear of the shuttlebox light CS and/or shuttlebox grid. This in turn could reduce the motivation for responding in the shuttlebox sufficiently to account for the more rapid rate of extinction in the irrelevant flood group as compared to the control groups. By this hypothesis, the effect of the irrelevant flooding treatment should be somewhat smaller than that of the relevant flooding treatment, and this effect was found. Extinction across CS modalities is not expected to be as complete as extinction of the CR that has actually undergone the extinction procedure (Konorski, 1948). Mineka reasoned that if this hypothesis were correct, then any fear extinction procedure for one CS should also hasten the extinction of the shuttlebox avoidance response, that is, flooding of an irrelevant avoidance CS would not be necessary to see the effect.

In a second experiment, Mineka (1976) found evidence to support this hypothesis. Two groups of rats were trained to perform a shuttlebox avoidance response and were given Pavlovian fear conditioning trials with a different CS in a jump-up box. One group also received fear extinction trials in the jump-up box, whereas the second group was returned to their home cages for a comparable amount of time. Both groups were then tested for extinction of the shuttlebox avoidance response. The fear conditioning and extinction group showed substantially faster extinction of the shuttlebox avoidance response than did the fear conditioning alone control group. So it seems that an irrelevant Pavlovian fear extinction process can mediate the extinction of the shuttlebox avoidance response. This suggests that the Pavlovian fear extinction that occurred during irrelevant flooding in the previous experiment may well be the factor that mediated shuttlebox extinction there also. Mineka (1976) concluded that these experiments suggest that generalization of extinction across CS modalities can mediate extinction of a shuttlebox avoidance response.

However, this conclusion that generalization of fear extinction across CS modalities mediated the irrelevant flooding effect is now questionable; more recent evidence from Mineka's laboratory described earlier (Mineka & Gino, 1979a) suggests that fear extinction does not mediate the relevant flooding effect. This raises the question of what is mediating the irrelevant flooding effect. Mineka (1974, 1976) rejected on a variety of empirical and theoretical grounds the competing-response, relaxation, and cognitive accounts of flooding as explanations for this effect. Crawford's (1977) SSDR hypothesis, however, provides a possible explanation of this irrelevant flooding effect that is worthy of investigation. This hypothesis would state that the irrelevant flooding procedure acts in the same way as does relevant flooding or confinement in fearful places—by producing a change in the dominant SSDR from fleeing to freezing.

There is one major weakness in the SSDR hypothesis that must be investigated before its feasibility can be determined. Mineka (1976) found in a third experiment that irrelevant flooding did not hasten jump-up box extinction, that is, there is an asymmetry to her irrelevant flooding effect. Yet the SSDR hypothesis originated from results with jump-up extinction, so irrelevant flooding should hasten jump-up extinction. There are two possible resolutions of this apparent paradox. First, the irrelevant flooding experiments used a better trained jump-up response than did the brief confinement experiments, and the confinement effect may only occur with less well-learned responses. Second, the brief confinement experiments used a different extinction procedure than did the irrelevant flooding experiment. In the former experiments the ani-

mals started extinction on the grid floor, and if they had any tendency to freeze, there was nothing to break up that tendency; by continuing to freeze they could rapidly meet the extinction criterion. Mineka, by contrast, used Baum's (1970a, 1970b) procedure, in which the animals are dumped off the safety ledge at the start of the first extinction trial. This dumping procedure may break up any tendency for freezing (perhaps it is even a mild punishment for freezing.) Hence any change in SSDR hierarchy that may have occurred during flooding or confinement in novel places might be reversed. The feasibility of these explanations can be determined by experiments designed to test whether the brief confinement effect occurs with the Baum/Mineka (ledge) extinction procedure and whether Mineka's irrelevant flooding procedure hastens jump-up extinction if the grids extinction procedure is used. (See Mineka & Gino, 1979b, for results which indicate that the effect of confinement in novel or fearful places does not occur either when the ledge extinction procedure is used or when the grids extinction procedure is used with a better learned response such as that used by Mineka, 1976).

In sum, there are several aspects to the nonspecificity of flooding, but the implications of findings in this area for our understanding of what goes on during traditional flooding procedures are as yet unclear. First, confinement in novel, fearful, or stressful places can produce floodinglike effects, but the extent to which these nonassociative effects contribute to traditional response prevention effects is unknown. It is unlikely, given Baum and LeClerc's (1974) time delay results as well as the results of Mineka and Gino (1979b), that purely nonassociative changes account for all of the effects of response prevention. Second, the effects of flooding one avoidance response can have transsituational effects and hasten the extinction of a second response that was learned to a different CS, but it is not yet known whether this irrelevant flooding effect is mediated by associative factors (e.g., generalization of extinction across CS modalities) or by nonassociative factors such as those suggested by Crawford's (1977) hypothesis.

The Role of Fear in Alternative Extinction Procedures for Avoidance

In recent years the conventional extinction procedure used to study the persistence of avoidance responses has received a substantial amount of criticism. In studies of the extinction of appetitive responses, the positive reinforcer—food or water—has traditionally been withheld. Analogously, it became common practice many years ago to withhold the negative reinforcer—shock—in studies of the extinction of avoidance behavior. Thus during extinction, shocks are no longer presented if an animal fails to respond, but responses continue to produce CS termination. Because it is now accepted that either CS termination and/or shock avoidance per se maintain avoidance behavior (Katzev, 1972; Herrnstein, 1969; Mackintosh, 1974), it seems to some that extinction of avoidance should be studied either by removing the CS termination contingency and/or by removing the shock avoidance contingency. Davenport and Olson (1968) reported rapid extinction of a discriminative bar-press avoidance response when both these sources of reinforcement were removed. Reynierse and Rizley (1970) subsequently confirmed these results with a shuttlebox avoidance response; removing the CS termination contingency or presenting CSs and USs randomly were both more effective procedures for eliminating avoidance than was the conventional procedure. Bolles, Moot, and Grossen (1971) further parceled out the relative contributions of CS termination versus shock avoidance to the extinction of avoidance and concluded that the shock avoidance contingency was crucial in creating an avoidance response that is highly resistant to extinction. Animals receiving shocks either randomly during extinction or as punishment for avoidance responses showed rapid extinction of the avoidance response. CS termination proved to be important only as long as the responses continued to avoid shock. This last finding conflicts with the report of Katzev that CS termination is extremely important in maintaining avoidance responding at a high rate when an avoidance contingency is in effect. Katzev reports low rates of responding with delayed CS termination and shock avoidance; Bolles

et al., on the other hand, report relatively high rates of responding with delayed CS termination and shock avoidance.[2]

It is now clear that the extreme persistence of avoidance responses is largely a result of using the conventional extinction procedures in which both CS termination and shock avoidance continue to be sources of reinforcement for responding in extinction. Because maintenance of the response/reinforcement contingency is necessary for a response that is highly resistant to extinction, some theorists (e.g., Mackintosh, 1974) have concluded that the longstanding question wrestled over by learning theorists of how to explain the persistence of avoidance responses in the face of apparent nonreinforcement (i.e., no shock) is really a pseudoproblem. These arguments are well-founded insofar as they constitute a sound analysis of what the reinforcing events and response/reinforcement contingencies are in avoidance learning. Such arguments fail to recognize, however, that the conventional extinction procedure that produces such extreme persistence is still interesting in its own right.

Aronfreed (1968) has cogently argued that the process of a child's internalizing control over his/her own behavior closely parallels the learning and maintenance of avoidance responses as they have conventionally been studied in animals. He argues that the extreme resistance to extinction seen in animals' avoidance responses (e.g., Solomon et al., 1953) is analogous to the observation that behavioral changes that occur in the process of human socialization are also highly resistant to changes in external contingencies. Children and adults frequently persist in responses originally acquired under circumstances of aversive control even when the aversive outcomes would no longer occur. Aronfreed further contends that this is most likely to occur in cases in which a child never had a cognitive representation of the original contingencies and so is unlikely to become aware of a change in the contingencies. This is obviously parallel to the case in which an animal learns an avoidance response that persists when conventional extinction procedures are used because it never stays around long enough to sample or become aware of the new contingencies. The newer extinction procedures, discussed previously, which change the CS termination and/or shock avoidance contingencies, forcibly expose the animal to the change in contingencies; hence extinction proceeds at a much faster rate.

Therefore, we can conclude that the extreme persistence of avoidance responses may not be as surprising or puzzling as it was once conceived to be and yet still concede that the question of what learning processes underlie the ultimate extinction of these responses is still an important and interesting one. If we are interested in the natural course of extinction, given the original contingencies, the processes underlying flooding or response prevention techniques, which hasten extinction, also remain important topics of investigation. These techniques are particularly interesting in light of the parallel dissociations between fear and avoidance behavior, on the one hand, and between subjective, behavioral, and psychophysiological indices of fear in humans undergoing systematic desensitization or flooding therapy, on the other hand.

One interesting issue that has not yet been investigated is the extent to which the dissociation between fear and avoidance that is observed following flooding and conventional extinction procedures also occurs with the newer procedures that remove the CS termination and/or shock avoidance contingencies. Some evidence (e.g., Coulter et al., 1969; Linton et al., 1970) suggests that the learning processes underlying flooding are different than those underlying the conventional extinction procedures because different residual amounts of fear remain. Other evidence (e.g., Berman & Katzev, 1972), however, suggests that flooding and regular extinction act via the same underlying process—nonreinforced CS exposure, which produces fear extinction. Riccio and Silvestri (1973) have speculated that extinction procedures that remove the

---

[2] It should be noted that punishment of avoidance often results in increased rather than reduced resistance to extinction—the so-called vicious circle phenomenon. (For extensive reviews of the literature on the determinants of which result is likely to occur in a particular situation, see, e.g., Brown, 1969, and Mackintosh, 1974.)

response/CS termination contingency may be more effective in eliminating the "motivational" (i.e., fear-eliciting) properties of the CS than are conventional extinction or response-blocking procedures. There is no compelling evidence to support or refute this speculation at present. It is unlikely that removing the shock avoidance contingency (e.g., Davenport & Olson, 1968) could produce extinction by fear extinction. This procedure is more likely to produce its effect through learning of a competing response or by a change in response/outcome expectancies. Recently, in fact, Overmier and Brackbill (1977) demonstrated that an effective avoidance extinction procedure involving noncontingent US presentations leaves fear of the CS intact. They interpret their results in terms of the independence of CS-fear and fear-avoidance response links in the avoidance-learning situation (see also Overmier & Bull, 1969). The results can also be interpreted in terms of the dissociation between fear and avoidance responding discussed earlier. Further exploration of these different avoidance procedures is particularly important because learning theorists and behavior therapy researchers are beginning to speculate, on the basis of animal experiments, about which contingencies it is most important to remove during desensitization and flooding therapy in humans (e.g., Riccio & Silvestri, 1973; Wilson, 1973).

In sum, we can see that there are multiple ways of extinguishing an avoidance response. What factors these procedures have in common is not yet clear, although it seems certain that multiple learning processes are involved. Experiments designed to determine the degree of residual fear following these different procedures will be of particular interest.

Conclusion

Fear and fear extinction do not appear to play any simple role in avoidance acquisition, maintenance, and extinction. There is often a marked dissociation or desynchrony between fear and avoidance behavior, and the determinants of this dissociation are as yet poorly understood. Fear attenuates over the course of avoidance learning; yet this cannot be explained by simple Pavlovian fear extinction (Starr & Mineka, 1977). The question of what continues to motivate avoidance responding as fear diminishes remains a hotly debated issue (e.g., D'Amato, 1970; Herrnstein, 1969; Mackintosh, 1974; Seligman & Johnston, 1973). These new theories of avoidance have emerged in large part as a result of the inadequacies of traditional two-process theory in explaining the dissociation between fear and avoidance behavior. Each theory postulates some new mechanism to explain well-maintained avoidance responding and extinction of responding. In so doing, these theories have tended to ignore or treat in only a cursory fashion the interesting determinants of this dissociation. However, fear remains a salient feature of avoidance learning, and the question of what determines its course at various stages of avoidance acquisition, maintenance, and extinction remains one of great practical and theoretical importance, at least for those interested in general questions about the determinants of fear and fear extinction in more complex situations than those traditionally studied by theorists of classical conditioning.

The determinants of fear extinction and its role in mediating avoidance extinction and the effects of response prevention techniques are also complex. Fear extinction is certainly not a necessary precursor of avoidance response extinction. With some avoidance responses (e.g., two-way shuttlebox), fear extinction may be sufficient to cause avoidance extinction, whereas with others, fear extinction may not be sufficient (e.g., one-way responses). Further work is needed to clarify the situations in which fear extinction is more or less central to avoidance response extinction.

The dissociation between fear and avoidance responding often observed following flooding is of particular interest because it may be functionally analogous to the dissociation frequently seen among different elements of the phobic response following systematic desensitization and flooding therapy (e.g., Lang, 1968, 1971; Rachman & Hodgson, 1974; Riccio & Silvestri, 1973). For example, patients frequently report that they can approach their phobic object (i.e., the

avoidance component is gone) but that they still feel afraid of it (i.e., the subjective feeling of fear and sometimes the psychophysiological concomitants of this feeling are still present). Although conditioned avoidance responses are often considered to be a poor model for human phobias (e.g., Costello, 1970; Seligman, 1971), many of the same variables found to enhance the effectiveness of flooding are also found to be important in flooding therapy with humans. (See Baum, 1970b, and Marks, 1972, for reviews.) So further work on the determinants of fear extinction during flooding is important for our understanding not only of the role of fear in the extinction of avoidance but also of a number of analogous problems in human behavior therapy techniques.

Reference Note

1. Mineka, S., & Gino, A. *Dissociation between CER and extended avoidance performance.* Manuscript submitted for publication, 1979.

References

Annau, Z., & Kamin, L. J. The conditioned emotional response as a function of intensity of the US. *Journal of Comparative and Physiological Psychology,* 1961, *54,* 428–432.

Aronfreed, J. *Conduct and conscience.* New York: Academic Press, 1968.

Baum, M. Effect of a loud buzzer applied during response prevention (flooding) in rats. *Behaviour Research and Therapy,* 1970, *8,* 287–292. (a)

Baum, M. Extinction of avoidance responding through response prevention (flooding). *Psychological Bulletin,* 1970, *74,* 276–284. (b)

Baum, M. Extinction of an avoidance response in rats via response prevention (flooding): A test for residual fear. *Psychological Reports,* 1971, *28,* 203–208.

Baum, M., & LeClerc, R. Irrelevant stress vs. response prevention (flooding) interpolated between avoidance acquisition and extinction in rats. *Journal of Psychiatric Research,* 1974, *10,* 307–314.

Benline, T. A., & Simmel, E. C. Effects of blocking of the avoidance response on the elimination of the conditioned fear response. *Psychonomic Science,* 1967, *8,* 357–358.

Berman, J. S., & Katzev, R. D. Factors involved in the rapid elimination of avoidance behavior. *Behaviour Research and Therapy,* 1972, *10,* 247–256.

Bersh, P. J., & Keltz, J. R. Pavlovian reconditioning and the recovery of avoidance behavior in rats after extinction with response prevention. *Journal of Comparative and Physiological Psychology,* 1971, *76,* 262–266.

Bersh, P., & Miller, S. K. The influence of shock during response prevention upon resistance to extinction of an avoidance response. *Animal Learning and Behavior,* 1975, *3,* 140–142.

Bersh, P. J., & Paynter, W. E. Pavlovian extinction in rats during avoidance response prevention. *Journal of Comparative and Physiological Psychology,* 1972, *78,* 255–259.

Black, A. H. The extinction of avoidance responses under curare. *Journal of Comparative and Physiological Psychology,* 1958, *51,* 519–524.

Black, A. H. Heart rate changes during avoidance learning in dogs. *Canadian Journal of Psychology,* 1959, *13,* 229–242.

Blackman, D. Conditioned suppression and the effects of classical conditioning on operant behavior. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Bolles, R. C. Species-specific defense reactions and avoidance learning. *Psychological Review,* 1970, *77,* 32–48.

Bolles, R. C. The avoidance learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 6). New York: Academic Press, 1972.

Bolles, R. C., Moot, S. A., & Grossen, N. E. The extinction of shuttlebox avoidance. *Learning and Motivation,* 1971, *2,* 324–333.

Boyd, T. L., & Levis, D. J. The effects of single component extinction of a three-component serial CS on the resistance to extinction of the conditioned avoidance response. *Learning and Motivation,* 1976, *7,* 517–531.

Brady, J. V. Experimental studies of psychophysiological responses to stressful situations. In *Symposium on medical aspects of stress in the military climate* (Walter Reed Army Institute of Research). Washington, D.C.: U.S. Government Printing Office, 1965.

Brady, J. V., & Harris, A. The experimental production of altered physiological states. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Brady, J. V., Kelly, D., & Plumlee, L. Autonomic and behavioral responses of the rhesus monkey to emotional conditioning. *Annals of the New York Academy of Science,* 1969, *159,* 959–975.

Brown, J. S. Factors influencing self-punitive locomotor behavior. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior.* New York: Appleton-Century-Crofts, 1969.

Brush, F. R. The effects of shock intensity on the acquisition and extinction of an avoidance response in dogs. *Journal of Comparative and Physiological Psychology,* 1957, *50,* 547–552.

Carlson, N. J., & Black, A. H. Traumatic avoidance learning: Note on the effect of response prevention during extinction. *Psychological Reports,* 1959, *5,* 409–412.

Coover, G. D., Ursin, H., & Levine, S. Plasma-corticosterone levels during active-avoidance learning in rats. *Journal of Comparative and Physiological Psychology*, 1973, *82*, 170–174.

Corriveau, D. P., & Smith, N. F. Fear reduction and "safety-test" behavior following response-prevention: A multivariate analysis. *Journal of Experimental Psychology: General*, 1978, *107*, 145–158.

Costello, C. G. Dissimilarities between conditioned avoidance responses and phobias. *Psychological Review*, 1970, *77*, 250–254.

Coulter, X., Riccio, D. C., & Page, H. A. Effects of blocking an instrumental avoidance response: Facilitated extinction but persistence of "fear." *Journal of Comparative and Physiological Psychology*, 1969, *68*, 377–381.

Crawford, M. Brief "response prevention" in a novel place can facilitate avoidance extinction. *Learning and Motivation*, 1977, *8*, 39–53.

D'Amato, M. R. *Experimental psychology: Methodology, psychophysics and learning*. New York: McGraw-Hill, 1970.

Davenport, D. G., & Olson, R. D. A reinterpretation of extinction in discriminated avoidance. *Psychonomic Science*, 1968, *13*, 5–6.

Davis, H. Conditioned suppression: A survey of the literature. *Psychonomic Monograph Supplements*, 1968, *2*(14, Whole No. 30). 283–291.

Denny, M. R. Relaxation theory and experiments. In F. R. Brush (Ed.), *Aversive conditioning and learning*. New York: Academic Press, 1971.

DeToledo, L., & Black, A. H. Heart rate: Changes during conditioned suppression in rats. *Science*, 1966, *152*, 1404–1406.

Dinsmoor, J. A. Punishment: I. The avoidance hypothesis. *Psychological Review*, 1954, *61*, 34–46.

Estes, W. K., & Skinner, B. F. Some quantitative properties of anxiety. *Journal of Experimental Psychology*, 1941, *29*, 390–400.

Gantt, W. H. Contributions to the physiology of the conditioned reflex. *Archives of Neurology and Psychiatry*, 1937, *37*, 848–858.

Gantt, W. H. Principles of nervous breakdown in schizokinesis and autokinesis. *Annals of the New York Academy of Sciences*, 1953, *56*, 141–165.

Grossen, N. E. Effect of aversive discriminative stimuli on appetitive behavior. *Journal of Experimental Psychology*, 1971, *88*, 90–94.

Hawk, G., & Riccio, D. C. The effect of a conditioned fear inhibitor (CS−) during response prevention upon extinction of an avoidance response. *Behaviour Research and Therapy*, 1977, *15*, 97–102.

Herrnstein, R. J. Method and theory in the study of avoidance. *Psychological Review*, 1969, *76*, 49–69.

Herrnstein, R. J., & Sidman, M. Avoidance conditioning as a factor in the effects of unavoidable shock on food-reinforced behavior. *Journal of Comparative and Physiological Psychology*, 1958, *51*, 380–385.

Hineline, P. N. Negative reinforcement and avoidance. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior*. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Hodgson, R., & Rachman, S. II. Desynchrony in measures of fear. *Behaviour Research and Therapy*, 1974, *12*, 319–326.

Hunt, H. F., & Brady, J. V. Some effects of electroconvulsive shock on a conditioned emotional response ("anxiety"). *Journal of Comparative and Physiological Psychology*, 1951, *44*, 88–98.

Irwin, F. W. *Intentional behavior and motivation: A cognitive theory*. Philadelphia, Pa.: Lippincott, 1971.

Kamin, L. J., Brimer, C. J., & Black, A. H. Conditioned suppression as a monitor of fear of the CS in the course of avoidance training. *Journal of Comparative and Physiological Psychology*, 1963, *56*, 497–501.

Katzev, R. What is both necessary and sufficient to maintain avoidance responding in the shuttlebox? *Quarterly Journal of Experimental Psychology*, 1972, *24*, 310–317.

Kimble, G. A. *Hilgard and Marquis' conditioning and learning*. New York: Appleton-Century-Crofts, 1961.

Konorski, J. *Conditioned reflexes and neuron organization*. Cambridge, England: Cambridge University Press, 1948.

Lang, P. J. Fear reduction and fear behavior: Problems in treating a construct. In J. M. Shlien (Ed.), *Research in psychotherapy* (Vol. 3). Washington, D.C.: American Psychological Association, 1968.

Lang, P. J. The application of psychophysiological methods to the study of psychotherapy and behavior modification. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.

Lederhendler, I., & Baum, M. Mechanical facilitation of the action of response prevention (flooding) in rats. *Behaviour Research and Therapy*, 1970, *8*, 43–48.

Levis, D. J., & Stampfl, T. G. Effects of serial CS presentation on shuttlebox avoidance responding. *Learning and Motivation*, 1972, *3*, 72–90.

Linden, D. R. Attenuation and reestablishment of the CER by discriminated avoidance conditioning in rats. *Journal of Comparative and Physiological Psychology*, 1969, *69*, 573–578.

Linton, J., Riccio, D. C., Rohrbaugh, M., & Page, H. A. The effects of blocking an instrumental avoidance response: Fear reduction or enhancement? *Behaviour Research and Therapy*, 1970, *8*, 267–272.

LoLordo, V. M. Positive conditioned reinforcement from aversive situations. *Psychological Bulletin*, 1969, *72*, 193–203.

LoLordo, V. M., & Rescorla, R. A. Protection of the fear-eliciting capacity of a stimulus from extinction. *Acta Biologiae Experimentalis*, 1966, *26*, 251–258.

Mackintosh, N. J. *The psychology of animal learning*. London: Academic Press, 1974.

Marks, I. Flooding (implosion) and allied treatment. In W. S. Agras (Ed.), *Behavior modifica-*

tion: Principles and clinical applications. Boston: Little, Brown, 1972.

Marrazo, M. J., Riccio, D. C., & Riley, J. Effects of Pavlovian conditioned stimulus-unconditioned stimulus pairings during avoidance response-prevention trials in rats. *Journal of Comparative and Physiological Psychology*, 1974, *86*, 96–100.

McAllister, W. R., & McAllister, D. E. Behavioral measurement of conditioned fear. In F. R. Brush (Ed.), *Aversive conditioning and learning*. New York: Academic Press, 1971.

Mineka, S. The effects of irrelevant flooding on the extinction of avoidance responses (Doctoral dissertation, University of Pennsylvania, 1974). *Dissertation Abstracts International*, 1974, *36*, 477B. (University Microfilms No. 75-14600)

Mineka, S. The effects of flooding an irrelevant response on the extinction of avoidance responses. *Journal of Experimental Psychology: Animal Behavior Processes*, 1976, *2*, 142–153.

Mineka, S., & Gino, A. Dissociative effects of different types and amounts of non-reinforced CS exposure on avoidance extinction and the CER. *Learning and Motivation*, 1979, *10*, 141–160. (a)

Mineka, S., & Gino, A. Some further tests of the brief confinement effect and the SSDR account of flooding. *Learning and Motivation*, 1979, *10*, 98–115. (b)

Monti, P. M., & Smith, N. F. Residual fear of the conditioned stimulus as a function of response prevention after avoidance or classical defensive conditioning in the rat. *Journal of Experimental Psychology: General*, 1976, *105*, 148–162.

Morokoff, P. J., & Timberlake, W. Cue exposure and overt fear responses as determinants of extinction of avoidance in rats. *Journal of Comparative and Physiological Psychology*, 1971, *77*, 432–438.

Morris, R. G. M. Pavlovian conditioned inhibition of fear during shuttlebox avoidance behavior. *Learning and Motivation*, 1974, *5*, 424–447.

Mowrer, O. H. On the dual nature of learning: A reinterpretation of "conditioning" and "problem-solving." *Harvard Educational Review*, 1947, *17*, 102–148.

Overmier, J. B., & Brackbill, R. M. On the independence of stimulus evocation of fear and fear evocation of responses. *Behaviour Research and Therapy*, 1977, *15*, 51–56.

Overmier, J. B., & Bull, J. A. On the independence of stimulus control of avoidance. *Journal of Experimental Psychology*, 1969, *79*, 464–467.

Page, H. A. The facilitation of experimental extinction by response prevention as a function of the acquisition of a new response. *Journal of Comparative and Physiological Psychology*, 1955, *48*, 14–16.

Pavlov, I. P. *Conditioned reflexes*. London: Oxford University Press, 1927.

Polin, A. T. The effect of flooding and physical suppression as extinction techniques on an anxiety-motivated avoidance locomotor response. *Journal of Psychology*, 1959, *47*, 253–255.

Rachlin, H., & Herrnstein, R. J. Hedonism revisited: On the negative law of effect. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior*. New York: Appleton-Century-Crofts, 1969.

Rachman, S. The passing of the two-stage theory of fear and avoidance: Fresh possibilities. *Behaviour Research and Therapy*, 1976, *14*, 126–131.

Rachman, S., & Hodgson, R. I. Synchrony and desynchrony in fear and avoidance. *Behaviour Research and Therapy*, 1974, *12*, 311–318.

Rescorla, R. A., & LoLordo, V. M. Inhibition of avoidance behavior. *Journal of Comparative and Physiological Psychology*, 1965, *59*, 406–412.

Rescorla, R. A., & Solomon, R. L. Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, 1967, *74*, 151–182.

Rescorla, R. A., & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. F. Prokasy (Eds.), *Classical conditioning II*. New York: Appleton-Century-Crofts, 1972.

Reynierse, J. H., & Rizley, R. C. Stimulus and response contingencies in extinction of avoidance by rats. *Journal of Comparative and Physiological Psychology*, 1970, *73*, 86–92.

Riccio, D. C., & Silvestri, R. Extinction of avoidance behavior and the problem of residual fear. *Behaviour Research and Therapy*, 1973, *11*, 1–9.

Schoenfeld, W. N. An experimental approach to anxiety, escape and avoidance behavior. In P. J. Hoch & J. Zubin (Eds.), *Anxiety*. New York: Grune & Stratton, 1950.

Scobie, S. R. Interaction of an aversive Pavlovian conditioned stimulus with aversively and appetitively motivated operants in rats. *Journal of Comparative and Physiological Psychology*, 1972, *79*, 171–188.

Seligman, M. E. P. Chronic fear produced by unpredictable electric shock. *Journal of Comparative and Physiological Psychology*, 1968, *66*, 402–411.

Seligman, M. E. P. Phobias and preparedness. *Behavior Therapy*, 1971, *2*, 307–320.

Seligman, M. E. P., & Johnston, J. A cognitive theory of avoidance learning. In F. J. McGuigan & D. B. Lumsden (Eds.), *Contemporary approaches to conditioning and learning*. New York: Wiley, 1973.

Seligman, M. E. P., Maier, S. F., & Solomon, R. L. Unpredictable and uncontrollable aversive events. In F. R. Brush (Ed.), *Aversive conditioning and learning*. New York: Academic Press, 1971.

Shearman, R. W. Response-contingent CS termination in the extinction of avoidance learning. *Behaviour Research and Therapy*, 1970, *8*, 227–239.

Shipley, R. H. Extinction of conditioned fear in rats as a function of several parameters of CS

exposure. *Journal of Comparative and Physiological Psychology,* 1974, *87,* 699–707.

Shipley, R. H., Mock, L. A., & Levis, D. J. Effects of several response prevention procedures on activity, avoidance responding, and conditioned fear in rats. *Journal of Comparative and Psychological Psychology,* 1971, *77,* 256–270.

Sidman, M. Avoidance conditioning with brief shock and no exteroceptive warning signal. *Science,* 1953, *46,* 253–261.

Sidman, M. Some properties of the warning stimulus in avoidance behavior. *Journal of Comparative and Physiological Psychology,* 1955, *48,* 444–450.

Sidman, M. By-products of aversive control. *Journal of the Experimental Analysis of Behavior,* 1958, *1,* 265–280.

Sidman, M., Herrnstein, R. J., & Conrad, D. G. Maintenance of avoidance behavior by unavoidable shocks. *Journal of Comparative and Physiological Psychology,* 1957, *50,* 553–557.

Solomon, R. L., Kamin, L. J., & Wynne, L. C. Traumatic avoidance learning: The outcomes of several extinction procedures with dogs. *Journal of Abnormal and Social Psychology,* 1953, *48,* 291–302.

Solomon, R. L., & Wynne, L. C. Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychological Review,* 1954, *61,* 353–385.

Soltysik, S. Inhibitory feedback in avoidance conditioning. *Boletin del Instituto des Estudios Medicos y Biologicos,* 1963, *21,* 433–449.

Stampfl, T. G., & Levis, D. J. *Implosive therapy: Theory and technique.* Morristown, N.J.: General Learning Press, 1973.

Starr, M. D., & Mineka, S. Determinants of fear over the course of avoidance learning. *Learning and Motivation,* 1977, *8,* 332–350.

Turner, L. H., & Solomon, R. L. Human traumatic avoidance learning: Theory and experiments on the operant-respondent distinction and failure to learn. *Psychological Monographs,* 1962, *76*(40, Whole No. 559).

Watson, J. P., & Marks, I. M. Relevant and irrelevant fear in flooding—A crossover study of phobic patients. *Behavior Therapy,* 1971, *2,* 275–293.

Weisman, R. G., & Litner, J. S. Positive conditioned reinforcement of Sidman avoidance behavior in rats. *Journal of Comparative and Physiological Psychology,* 1969, *68,* 397–603.

Weisman, R. G., & Litner, J. S. The role of Pavlovian events in avoidance training. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and learning.* New York: Academic Press, 1972.

Werboff, J., Duane, D., & Cohen, B. D. Extinction of conditioned avoidance and heart rate responses in rats. *Journal of Psychosomatic Research,* 1964, *8,* 29–33.

Wilson, G. T. Counterconditioning versus forced exposure in extinction of avoidance responding and conditioned fear in rats. *Journal of Comparative and Physiological Psychology,* 1973, *82,* 105–114.

Wolpe, J. *Psychotherapy by reciprocal inhibition.* Stanford, Calif.: Stanford University Press, 1958.

Received April 25, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1011–1014

Testing for Association in 2 × 2 Contingency Tables With Very Small Sample Sizes

Gregory Camilli and Kenneth D. Hopkins
Laboratory of Educational Research
University of Colorado

Applied statistics textbooks generally recommend the use of the chi-square tests of homogeneity and independence with 2 × 2 contingency tables only when the expected frequency of each cell is five or more. Recent research has shown this rule-of-thumb criterion to be unnecessarily restrictive, but has not explored the accuracy of the chi-square tests when the total number of observations is less than 20 or when the expected frequencies fall well below one—the primary issues considered in this article. The chi-square tests of homogeneity and independence were found to provide reasonably accurate estimates of Type I error probability for $N \geq 8$. Certain alternatives to the chi-square tests are considered.

There are three distinct models for deriving probability statements for 2 × 2 contingency tables (Kendall & Stuart, 1967, pp. 549–555). In Model 1, a researcher specifies both sets of marginal frequencies before the data are collected; this is Fisher's exact test. In Model 2, one set of marginal frequencies is fixed, the other being free to vary randomly; this is termed a test of *homogeneity*. In Model 3, neither set of marginal frequencies is fixed; this is typically described as a test of *independence*. The chi-square test has been shown to provide approximate, but accurate, probability estimates in Case 2 (test of homogeneity) and Case 3 (test of independence) for sample sizes as small as $N = 20$ if the Yates correction is not applied (Camilli & Hopkins, 1978; Roscoe & Byars, 1971). The accuracy of chi-square tests of homogeneity and independence with sample sizes smaller than 20 does not appear to have been explored. Although Model 1 is rarely consonant with empirical data, the chi-square test in which the continuity correction has been made estimates probabilities associated with Model 1. In Model 2 or Model 3 chi-square applications, the continuity correction should not be applied, since it decreases the accuracy of the related probability statements (Camilli & Hopkins, 1978).

Requests for reprints should be sent to Kenneth D. Hopkins, Laboratory of Educational Research, University of Colorado, Boulder, Colorado 80309.

Method

This study examined the effects of the four independent variables listed below on the actual proportion of Type I error in relation to the nominal alpha values of .10, .05, and .01: (a) total number of observations, N, in the 2 × 2 contingency table (N = 4, 8, 12, 16, and 20); (b) row marginals fixed (Model 2) versus random (Model 3); (c) relative frequencies of the N observations in the two row categories when row marginal frequencies ($n_{1\cdot}$ and $n_{2\cdot}$) were fixed ($n_{1\cdot}/n_{2\cdot}$ = 1.0 and .33); when row frequencies were random, the row marginal proportions varied randomly about the parameters $\pi_{1\cdot}$ and $\pi_{2\cdot}$, with $\pi_{1\cdot} = 1 - \pi_{\cdot 2}$ = .500, .707, and .794; and (d) proportion of N observations in the two random column categories $\pi_{\cdot 1}$ and $\pi_{\cdot 2}$ ($\pi_{\cdot 1} = 1 - \pi_{\cdot 2}$ = .500, .707, and .794). (The proportions employed by Roscoe & Byars, 1971, and Camilli & Hopkins, 1978, were used in this study to facilitate comparisons of findings.)

Results

Model 2: Chi-Square Test of Homogeneity

The actual proportions of the 10,000 replications in which the observed chi-square value exceeded the critical chi-square value at nominal alpha levels of .10, .05, and .01 are given in the upper portion of Table 1 for selected Model 2 chi-square applications. For example, the second row of Table 1 indicates

Table 1
Actual Proportions of Type I Errors in 10,000 Replications Given by Chi-Square Tests of Homogeneity (Model 2) and Independence (Model 3) in Selected 2 × 2 Contingency Tables [a]

Random column proportions			Row frequency/ proportion		Proportion of Type I errors ($\hat{\alpha}$)[b]			Uncomputable[c]
$\pi_{.1}$	$\pi_{.2}$	N	$n_{1.}$	$n_{2.}$	.10	.05	.01	χ^2
			Model 2					
.5	.5	4	2	2	.1224[d]	.1224[d]	.0000	1294
		8	4	4	.0715	.0715	.0058	74
		12	6	6	.1426	.0576	.0066	3
		16	8	8	.0746	.0703	.0048	1
		20	10	10	.1160	.0400	.0132	0
.794	.206	4	1	3	.1172	.1172	.0000	3920
		8	2	6	.1069	.0274	.0110	1631
		12	3	9	.1076	.0222	.0178	611
		16	4	12	.1052	.0318	.0228	253
		20	5	15	.1088	.0358	.0188	100
			Model 3					
.5	.5	4	$\pi_{1.} = .5$	$\pi_{2.} = .5$	.1077	.1077	.0000	2359
		8			.1084	.0656	.0082	173
		12			.1527	.0572	.0066	12
		16			.1114	.0708	.0083	1
		20			.1203	.0507	.0117	0
.5	.5	4	.794	.206	.0759	.0759	.0000	4722
		8			.0684	.0319	.0062	1602
		12			.1006	.0390	.0047	613
		16			.1063	.0411	.0057	231
		20			.1048	.0416	.0061	94
.794	.206	4	.794	.206	.0570	.0570	.0000	6342
		8			.0824	.0303	.0161	2913
		12			.0896	.0403	.0117	1221
		16			.0967	.0435	.0183	482
		20			.0928	.0483	.0143	205

[a] Pseudorandom numbers were generated by the International Mathematical and Statistical Libraries subroutine GGUB (IMSL, 1977). In addition to the 25 simulations reported in Table 1, 25 additional simulations were performed with row and column parameters falling between the extremes represented in Table 1.

[b] $\sigma_p = .0030$, .0022, and .0010 for $\alpha = .10$, .05, and .01, respectively.

[c] In certain instances of the 10,000 replications in which the expected cell frequencies were very small, the random number generating process yielded a column with a zero frequency, hence the chi-square was not computable. One can subtract this figure from 10,000 to recalculate the proportion of Type I errors in the sample space in which the chi-square was computable.

[d] Because of the discreteness in the empirical sampling distributions, the proportions of Type I errors are identical at $\alpha = .10$ and $\alpha = .05$ (and in all other simulations in which $N = 4$).

that when the 2 × 2 contingency table contained $N = 8$ observations divided equally between the two rows ($n_{1.} = n_{2.} = 4$) and when the proportion of N observations falling in each column varied randomly about $.5 = \pi_{.1} = \pi_{.2}$, the observed proportions of Type I errors ($\hat{\alpha}$) were .0715, .0715, and .0058 at the nominal alpha values of .10, .05, and .01, respectively.

Model 3: Chi-Square Test for Independence

Selected results from applications of Pearson's chi-square to null situations in which both row and column proportions differed randomly from the specified parameters $\pi_{1.}$ and $\pi_{.1}$ are given in the lower portion of Table 1. In all 50 simulations (25 of which are reported in Table 1), the expected frequencies

of at least two of the four cells were five or less; the expected frequency of one or more cells was extremely small (one or less) in 24 of the simulations.

For both models, when $N \geq 8$, the nominal alpha values were reasonably accurate at $\alpha = .05$ (the proportion of Type I errors varied between .0222 and .0715) and $\alpha = .01$ (the proportion of Type I errors varied between .0047 and .0228). For $N = 4$, the accuracy at $\alpha = .10$ was adequate, but for $\alpha = .05$ and $\alpha = .01$, accuracy was very poor. The issue of power is critically important. A later section demonstrates that when the two factors are very highly correlated, the statistical power of the chi-square test with very small Ns is very low. But alternatives can provide a more powerful test under certain conditions.

Alternatives to 2 × 2 Chi-Square Tests

To illustrate the first alternative, consider a 2 × 2 contingency table in which $N = 10$ and $n_{11} = 4$, $n_{12} = 1$, $n_{21} = 1$, and $n_{22} = 4$, hence $n_{1\cdot} = n_{2\cdot} = n_{\cdot 1} = n_{\cdot 2} = 5$. The probabilities for this particular configuration of frequencies for Models 2 and 3 are related to the probability of Model 1 by binomial multipliers (Kendall & Stuart, 1967, pp. 551–555):

$$p_1 = \frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{N!n_{11}!n_{21}!n_{12}!n_{22}!}; \qquad (1)$$

$$p_2 = p_1\binom{N}{n_{\cdot 1}}\pi_{\cdot 1}{}^{n_{\cdot 1}}(1 - \pi_{\cdot 1})^{N-n_{\cdot 1}}; \qquad (2)$$

$$p_3 = p_2\binom{N}{n_{1\cdot}}\pi_{1\cdot}{}^{n_{1\cdot}}(1 - \pi_{1\cdot})^{N-n_{1\cdot}}. \qquad (3)$$

Using Equation 1, p_1 (Fisher's exact probability) is found to be .0992. Using Equations 2 and 3 and supposing that $\hat{\pi}_{\cdot 1} = \hat{\pi}_{1\cdot} = .5$, the probabilities of the observed 2 × 2 configuration can be determined:

$$\hat{p}_2 = p_1\left(\frac{10!}{5!5!}\right)(.5)^{10} = .0992(.2461) = .0244$$

and

$$\hat{p}_3 = \hat{p}_2(.2461) = .0060.$$

But these probabilities are for observing this particular frequency configuration in the sample space. The random marginal frequencies allow many more possible frequency configurations under Model 2 than under Model 1 and many more still under Model 3. Indeed, unless N is very small (10 or less), the process of enumerating the various frequency configurations, and hence deriving a probability distribution, is not practicable without the assistance of computer programs. Furthermore, this "exact" distribution will only be as useful as $\hat{\pi}_{1\cdot}$ and $\hat{\pi}_{2\cdot}$ are accurate.

A modification of Fisher's exact test, proposed by Tocher (1950), can be used to compensate for the discreteness of sampling distributions if one desires to make decisions at precise conventional alpha values such as $\alpha = .05$. In Tocher's procedure, one selects alpha (α) and then calculates the cumulative Model 1 distribution for the given configuration: L_1 is defined as the probability of all more extreme configurations (in our example, $n_{11} = n_{22} = 5$ and $n_{12} = n_{21} = 0$), and $L_1 = .0040$ using Equation 1); L_2 is the probability of the obtained configuration (.09921) plus L_1, hence $L_2 = .1032$. If $L_2 \leq \alpha$, H_0 is rejected; if $L_1 > \alpha$, H_0 is tenable. But in situations like ours in which $L_2 > \alpha > L_1$, the ratio R is determined: $R = (\alpha - L_1)/(L_2 - L_1)$. In our example,

$$R = (.05 - .004)/(.1032 - .004) = .464.$$

A random number (X) is then selected from a uniform distribution within the interval (0, 1). If $X \leq R$, H_0 is rejected; otherwise H_0 remains tenable; or given our data set, the probability that H_0 will be rejected is .464. The exact level of significance equals $XL_2 + (1 - X)L_1$.

Tocher's test (also known as the randomized exact test) has been shown to be the uniformly most powerful unbiased test for all three models (Kendall & Stuart, 1967, p. 554; Tocher, 1950). The degree of difference in power does not seem to have been explored for very small Ns.

Power Comparisons

The relative power of the chi-square and Tocher's exact test was compared for 12 Model 2 and Model 3 situations that were a subset of the original 50 simulations, except that the cell proportions (π_{ij}) varied from .4 to .1. At $\alpha = .05$ and at $N = 8$, the power efficiency of the chi-square ranged from .18 to

.64, with the lesser power differences occurring with larger $\pi_{11}/\pi_{1\cdot} - \pi_{21}/\pi_{2\cdot}$ values. The differences in power were much less for $12 \leq N \leq 20$, with the power efficiency of the chi-square ranging from .64 to 1.05. The gain from Tocher's test is negligible if $N \geq 20$ (Starmer, Grizzle, & Sen, 1974). The absolute power of the chi-square or the Tocher tests exceeded .80 only when effects were very large (e.g., $\pi_{11}/\pi_{1\cdot} = .8$ and $\pi_{21}/\pi_{2\cdot} = .2$) and $N \geq 16$ with $\alpha = .05$. Researchers working with precious observations that may limit N should probably relax alpha to .10 and make directional tests to increase power.

Many behavioral researchers will find Tocher's test philosophically objectionable in spite of its greater power, since the gain in power is affected not by data but by the "luck of the draw."

Conclusions

In 2×2 contingency tables, the chi-square test provides a quick and reasonably accurate Type I probability statement for tests of homogeneity or tests of independence if $N \geq 8$. If $N < 5$, $\alpha = .10$ should be used, since $\alpha = .05$ and $\alpha = .01$ become very inaccurate.

Fisher's exact test can also be employed. but is very conservative for Model 2 and Model 3 applications (Camilli & Hopkins, 1978). The most powerful, but philosophically controversial, alternative is Tocher's exact test, which has greater power than the chi-square test, especially when $N < 20$. To increase power if N cannot be increased, researchers should consider relaxing alpha to .10 or more and making one-tailed tests.

References

Camilli, G., & Hopkins, K. D. Applicability of chi-square to 2×2 contingency tables with small expected cell frequencies. *Psychological Bulletin*, 1978, *85*, 163–167.

International Mathematical and Statistical Libraries. *Library 3 reference manual* (Vol. 1, 6th ed.). Houston, Tex.: Author, 1977.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 2, 2nd ed.). New York: Hafner, 1967.

Roscoe, J. T., & Byars, J. A. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association*, 1971, *66*, 755–759.

Starmer, C. F., Grizzle, J. E., & Sen, P. K. Some reasons for not using the Yates continuity correction on 2×2 contingency tables: Comment. *Journal of the American Statistical Association*, 1974, *69*, 376–378.

Tocher, K. D. Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika*, 1950, *37*, 130–144.

Received May 1, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1015–1043

Psychological Control of Essential Hypertension: Review of the Literature and Methodological Critique

Peter Seer
Department of Psychiatry, University of Auckland, Auckland, New Zealand

Recent studies (1971–1978) that investigated psychological approaches to the treatment of essential hypertension are reviewed. Twenty studies that use techniques of biofeedback, relaxation, and meditation training are summarized in table form. They are subjected to a detailed methodological critique, and suggestions for methodological improvements and directions for future research are proposed. Most experiments demonstrated blood pressure reductions too small to be of clinical significance. A combination of biofeedback and relaxation/meditation with other behavioral techniques appears most promising, and suggestions for a more comprehensive approach to assessment and training are made. Although studies comparing biofeedback and relaxation/meditation were inconclusive, relaxation/meditation is suggested to hold more promise because it requires no sophisticated technology and has been reported to simultaneously reduce other stress-related complaints.

In recent years we have witnessed an increasing interest in the psychological control of essential hypertension and of disease in general. This trend appears to be related to a shift in disease patterns. On the one hand, we can observe a drastic reduction in contagious diseases and, on the other hand, a dramatic increase in degenerative disorders such as coronary heart disease and essential hypertension (Stoyva, 1976). There is an extensive body of research indicating that essential hypertension can be related to dysfunctional habits of living and of responding to an environment of ever increasing complexity (Gutmann & Benson, 1971; Henry & Cassel, 1969). Further related to this trend is the growing recognition that the pharmacological treatment of essential hypertension has many undesirable side effects (Bulpitt & Dollery, 1973) and is far less effective than drug advertisements lead one to believe (Kannel & Dawber, 1973; LoGerfo, 1975). The development and scientific study of nondrug alternatives in the control of essential hypertension is therefore highly desirable. Psychological approaches such as biofeedback, relaxation, and meditation constitute such an alternative. The purpose of this article is to present a review of recent research in this field (1971–1978), to critically evaluate its methodology, to make special recommendations for minimal methodological requirements, and to make suggestions for future research.

Definition, Incidence, and Classification of Essential Hypertension

Hypertension has become an epidemic of major proportions in western society. Estimates of its prevalence vary from 10% to as high as 30% of the total adult population, depending on the definition of what constitutes high blood pressure (Stamler, Stamler, Riedlinger, Algera, & Roberts, 1976). Although there exists considerable disagreement over the definition of hypertension, it is safe to say that for persons of up to 50 years of age, a blood pressure of 145/95 mmHg would be classified as mild hypertension. Hyperten-

I wish to thank John Raeburn and John Spicer for their critical reading of earlier drafts of the manuscript.

Requests for reprints should be sent to Peter Seer, who is now at Fachbereich Psychologie, Rehabilitationszentrum, Südring 15, 7812 Bad Krozingen, West Germany.

sion in its early stages is asymptomatic, that is, it is not accompanied by any overt warning signs. Consequently, as many as 50% of all cases of hypertension go undetected (Onesti, Kim, & Moyer, 1973). More than 90% of all cases of hypertension are of unknown etiology; they fall into the category of primary or essential hypertension. The remainder, labeled secondary hypertension, is due to identifiable renal, endocrine, neurogenic, and other disorders; this latter category of secondary hypertension will not be discussed in this article.

Hypertension as a Major Risk Factor

It has been firmly established that moderate to severe hypertension increases morbidity as well as mortality from many diseases, especially cardiovascular disorders. The most comprehensive and conclusive epidemiological study in this field is the Framingham study (Kannel, 1976), which has been in continuous operation since 1948. This study has found hypertension to be one of the most robust predictors of such life-threatening disorders as myocardial infarction, congestive heart failure, stroke, and damage to kidneys, eyes, and other organs. Risk was found to increase proportional to increases in blood pressure, and hypertensives were found to be three times more likely to develop cardiovascular disease than normotensives. Both systolic and diastolic blood pressure were found to be of equal importance.

If we consider that, for example, in the United States more than 50% of all deaths are linked to cardiovascular and cerebrovascular disorders, and if we consider the enormous suffering of those affected and the staggering costs involved, it becomes obvious that prevention and treatment of hypertension and other known risk factors is of paramount importance.

Consequences of Reducing Elevated Blood Pressure

The importance of controlling moderate to severe hypertension for the reduction of morbidity and the prolongation of life has been most convincingly demonstrated in two major double-blind studies by the Veterans Administration Cooperative Study Group on Antihypertensive Agents (1967, 1970). Results clearly indicated that in comparison to experimental subjects who received active pharmacological treatment, a significantly higher percentage of placebo group subjects developed congestive heart failure, stroke, renal damage, retinopathy, and accelerated hypertension. However, pharmacological treatment did not significantly reduce the incidence of myocardial infarction.

Medical opinion differs considerably with regard to the value of the treatment of mild hypertension, in the absence of clear evidence of its general preventative worth. However, the control of even small blood pressure elevations has been shown to reduce cardiovascular disease in subjects who simultaneously display other cardiovascular risk factors such as excessive weight, smoking, and elevated cholesterol levels (Kannel, 1976).

Psychophysiological View of Essential Hypertension

Despite some 40 years of intensive research into the mechanisms and causes of essential hypertension, its etiology is still unknown. A number of physical correlates such as hereditary predisposition, salt intake, overweight, and abnormalities in the renin-angiotensin system (Kannel & Dawber, 1973) have been isolated, but researchers disagree widely as to their respective importance.

Temporary rises in blood pressure in response to events that are subjectively perceived as exciting, demanding, or distressing have been observed. Essential hypertension has been found in people whose adaptive capabilities had been overtaxed in situations such as natural disaster or war, hazardous work environments and excessive work pressures, job loss and unemployment, migration and urbanization, and others. Further, research has shown that sustained blood pressure elevations can be produced experimentally in animals by prolonged elicitation of the emergency reaction, using such procedures as electrical stimulation of the hypothalamus, exposure to various aversive stim-

uli, classical and operant conditioning, and disruption of normal social interrelationships. These blood pressure elevations were found to persist even after the termination of the aversive event. The role of psychosocial factors in essential hypertension has been reviewed by Gutmann and Benson (1971) and Henry and Cassel (1969).

The apparent role of psychosocial factors has led to the development of psychological approaches to the treatment of essential hypertension. These approaches are usually based on a psychophysiological model in which the repeated and prolonged elicitation of the "emergency reaction" (Cannon, 1932) with its characteristic blood pressure lability eventually leads to stabile hypertension in predisposed individuals (Gutmann & Benson, 1971; Henry & Cassell, 1969; Stoyva, 1976). Most studies attempting to control essential hypertension with the psychological techniques of relaxation/meditation and biofeedback aim at reducing the sympathetic nervous system activity that mediates the emergency reaction. It is important to stress that the exact mechanism linking transient rises in blood pressure with sustained elevations is still unknown and that other models can also be used to understand the disorder.

Description of Psychological Techniques Used in the Control of Essential Hypertension

The most widely researched psychological approaches to the control of essential hypertension fall into two major groups, namely, relaxation/meditation and biofeedback.

Relaxation/Meditation

Progressive relaxation. Among the approaches explicitly aimed at producing physical relaxation, one of the most widely used is Jacobson's (1970) "progressive relaxation." Here, the person is instructed to tense and relax various groups of striate muscles throughout the body, starting, for example, with the hands and arms and then progressing to facial muscles and muscles in the trunk and legs. It is worth mentioning that Jacobson (1939) was the first to observe decreases in blood pressure concomitant with muscular relaxation. Progressive relaxation is also an integral part of "metronome-conditioned relaxation," a technique developed by Brady (1973). Here, the person is first taught to tense and relax major muscle groups. This is followed by more general instructions to "re-lax and let go" paced by a metronome set at 60 beats/min.

Meditation techniques. Meditation techniques, as a rule, were conceived within a particular philosophical and religious context. Only recently have they been used outside this context or been combined with various behavior change methods (Shapiro & Zifferblatt, 1976). Meditation techniques are difficult to define as a whole; only the "concentrative" meditation techniques will be considered here. (For a detailed discussion of meditation from a psychological point of view, see Naranjo & Ornstein, 1971.) Their common feature seems to lie in learning to direct one's attention toward a "mental device" (Benson, 1975) or a focus with which the student becomes passively absorbed. A variety of mental devices such as mantras, chants, prayers, visual symbols, or one's breathing or heartbeat are used in the various meditation techniques. Their function is to reduce or eliminate conceptual thinking ("the mental chatter") and to facilitate the development of an encompassing focus on the present moment and concomitant feelings of calm and relaxation. A common phenomenon in both relaxation and meditation techniques is that the practitioner frequently finds his or her attention shifting away from the mental device toward unrelated thoughts, ideas, images, preoccupations, worries, or sensations. These task-irrelevant thought processes are usually not dealt with satisfactorily in the relaxation techniques, whereas the meditation approaches commonly contain explicit instructions for dealing with them.

One of the more common meditation techniques is attention to the ingoing and outgoing flow of one's breath without controlling it. This technique of breath meditation has been popularized and simplified for use with hypertensive patients by Benson (1975) and

has been applied in several hypertension studies. It requires the person to think of the number 1 (or count up to higher numbers) with each exhalation and to return to breath counting whenever distractions occur.

The techniques described so far all involve attending to physical sensations or processes (e.g., breathing). In transcendental meditation (TM; Mahesh Yogi, 1968) a mantra or meaningless sound is introduced as an attentional focus, and the person is taught to repeat the sound mentally in an effortless way.

Physiological effects and active components of relaxation/meditation techniques. A considerable amount of research data on the physiological effects of relaxation/meditation has been accumulated in recent years. Results clearly indicate that physiological effects are not consistent across different forms of relaxation, and meditation and may even differ within a given technique depending on such variables as mode of instruction, context within which the technique is taught, number and length of training sessions, and type of subject population (Woolfolk, 1975). Regarding this latter variable, Davidson and Schwartz (1976) have convincingly argued that different clinical problems manifest themselves in different modes and physiological systems. Therefore, they may require different relaxation/meditation techniques that specifically respond to the involved cognitive, somatic, or attentional processes.

However, there is considerable evidence that common to these techniques is the elicitation of a general physiological pattern that Stoyva and Budzynski (1974) termed "cultivated low arousal." It is characterized by decreases in pulse and respiratory rate, decreases in muscle tonus and oxygen consumption, and increases in skin resistance. This "relaxation response" (Benson, 1975) is seen as incompatible with and counteracting the emergency reaction. Its frequent elicitation is assumed to lead eventually to a reduction in sympathetic activity and consequently to a lowering of blood pressure.

Benson (1975) defined four components of the common active ingredients of relaxation/meditation: (a) a quiet environment, (b) a decreased muscle tonus, (c) a passive attitude, and (d) the restriction of one's attention to a mental device. This latter component is generally seen as most important in explaining the mechanisms that underlie relaxation/meditation (Naranjo & Ornstein, 1971). There is a great need for more research, as the only well-controlled study (Smith, 1976) found that neither a mental device nor a passive attitude were necessary components of meditation.

Common to all relaxation/meditation techniques is that they require the student to intersperse his or her ongoing activities once or twice daily with a 15–20-minute period of just sitting quietly. This component alone may sufficiently explain the observed changes. Several studies comparing the physiological short-term effects of TM with a control condition of just sitting found that they both resulted in comparable physiological changes (Travis, Kondo, & Knott, 1976; Treichel, Clinch, & Cran, 1973; Walrath & Hamilton, 1975).

Biofeedback

In contrast to the relaxation/meditation techniques, biofeedback requires a highly sophisticated technology that has been developed only in recent years. In blood pressure biofeedback training, pressure is recorded on a continuous or noncontinuous basis, and the information is fed back to the subject in the form of a light and/or sound signal or by letting subjects directly observe their blood pressure record. This allows the subject to become aware of fluctuations in blood pressure and to learn to exercise voluntary control.

Constant cuff technique. Shapiro and his associates (e.g., Shapiro, Tursky, Gershon, & Stern, 1969) have made the most important contribution in this field with their "constant cuff" method that allows continuous, beat-by-beat feedback of blood pressure. It involves mounting a crystal microphone inside a standard pressure cuff and placing it over the brachial artery. The cuff is inflated and set at a constant pressure close to the person's average systolic or diastolic blood pressure. Whenever the person's blood

pressure rises above this set level, a Korotkoff sound can be detected by the microphone. With each beat of the heart, the person receives information (yes–no feedback) as to whether his or her blood pressure is above or below the set average blood pressure (binary feedback). After a trial of 50 consecutive heartbeats, the cuff is deflated for 30 sec and then set at a new level depending on the average blood pressure in the previous trial. One session consists of 25 to 30 trials. Shapiro et al. (1969) and Shapiro, Schwartz, and Tursky (1972) were the first to show that small but reliable changes in systolic and diastolic blood pressure could be achieved with this method in normotensive volunteer subjects. (For a review of blood pressure biofeedback studies with normotensives, see Blanchard & Young, 1973.) Benson, Shapiro, Tursky, and Schwartz (1971) were the first to demonstrate the potential usefulness of blood pressure biofeedback training in the treatment of essential hypertension. Since then several other blood pressure biofeedback studies have been conducted, which are presented in the following section.

The constant cuff method has recently been further improved (Elder, Longacre, Welsh, & McAfee, 1977) by adding a tracking device that monitors blood pressure beat by beat and automatically adjust cuff pressure every three to four heartbeats. Two pressure cuffs, one on each arm, that are alternately inflated for 100 sec are used. Feedback is given via an audible tone that changes in pitch with fluctuations in blood pressure. So far this method has only been reported in experiments with normotensives (Elder, Welsh, Longacre, & McAfee, 1977) but is likely to be used with hypertensive subjects in the near future.

Pulse wave velocity technique. Pulse wave velocity as an indirect measure of mean arterial pressure has recently received considerable attention (e.g., Gribbin, Steptoe, & Sleight, 1976). It has also been used for feedback purposes with essential hypertensives (Walsh, Dale, & Anderson, 1977). In Walsh et al. pulse wave velocity was determined by measuring the pulse transit time between the heart's right ventricular action and the finger pulse. It was fed back to the subject both auditorily, in the form of a tone that became higher or lower as the pulse transit time increased or decreased, and visually, on an oscilloscope (analogue feedback). Pulse transit time has been found to relate inversely to mean arterial pressure (Steptoe, Smulyan, & Gribbin, 1976). Despite its great advantage of avoiding the intrusive effects of constant cuff pressure and repeated cuff inflations, this technique has so far not been applied on a larger scale with essential hypertensives.

Noncontinuous blood pressure feedback techniques. Two noncontinuous feedback techniques have been studied in the control of essential hypertension. Elder, Ruiz, Deabler, and Dillenkoffer (1973) applied a blood pressure recorder that automatically measured diastolic blood pressure every 2 min for a total of 20 successive determinations. Subjects received a red-light signal (and verbal praise) contingent on increasingly larger blood pressure reductions. Subjects in Shoemaker and Tasto's (1975) experiment were able to directly observe their blood pressure recording at 90-sec intervals via a mirror placed above the chart recorder. Two straight lines representing the person's pretrial average systolic and diastolic blood pressures were superimposed on the recording, and subjects were instructed to lower their blood pressure so that the recorded blood pressure would fall below the lines.

Noncontinuous blood pressure feedback has the disadvantage of not being sufficiently sensitive to cope with the inherent variability of blood pressure. Shannon, Goldman, and Lee (1978), comparing three types of systolic blood pressure feedback with normotensive subjects, found continuous (beat-by-beat) binary feedback clearly superior to a relatively continuous proportional and a noncontinuous (75-sec intervals) proportional feedback condition. This suggests that the greater the time lag between beat-by-beat blood pressure changes and feedback, the less effective is the technique.

Physiological effects and active components of blood pressure biofeedback. Al-

though a great number of experiments with normotensive subjects have been conducted to tease out the essential ingredients in blood pressure feedback, the exact mechanisms and processes involved are still unknown. Whereas the relaxation/meditation approaches aim at an indirect control of blood pressure via a generalized relaxation response, blood pressure biofeedback aims at specifically or directly altering blood pressure as such. But experimental data are inconsistent. Some researchers have confirmed this specificity of learned blood pressure control (e.g., Kristt & Engel, 1975), whereas others have reported concomitant changes in other cardiovascular parameters (e.g., Fey & Lindholm, 1975).

It is generally assumed that the beat-by-beat feedback of blood pressure is the most important element in the technique. However, there is experimental evidence that normotensive and hypertensive subjects can change blood pressure by instruction alone, that is, by simply being told to change blood pressure in the desired direction without instructions on how to achieve this. Redmond, Gaylor, McDonald, and Shapiro (1974), using essential hypertensives as subjects, found instructions alone to be effective in reducing systolic and diastolic blood pressure by 8–14 mmHg and 6–11 mmHg, respectively. Results with normotensives have been inconsistent. In Steptoe's (1976) experiment, normotensive subjects who received instructions alone reduced blood pressure equally as effectively as subjects who, in addition, received pulse wave velocity feedback. The addition of exteroceptive feedback, however, enhanced the increase condition. In a subsequent study, Steptoe (1977) controlled for environmental stimulation by exposing subjects in the instruction-only group to identical visual displays as feedback subjects. Results favored the feedback condition that produced greater increases in pulse transit time and by inference, greater decreases in mean arterial blood pressure. Much more could be said about this issue, but it would be beyond the scope of this article. (For a detailed discussion, see Brener, 1974, and Shapiro, 1977.)

More research is necessary to determine the active ingredients in blood pressure biofeedback training. The role of attentional, cognitive, and imagery processes as well as the role of reinforcement and individual differences is far from clear. To develop the most effective blood pressure technique, studies are needed that compare the constant cuff and the pulse velocity techniques, continuous and noncontinuous, binary and proportional feedback, and systolic and diastolic blood pressure feedback.

Electromyograph (EMG) and Galvanic Skin Response (GSR) feedback. EMG and GSR feedback, alone or in combination with other techniques, have also been studied in the control of essential hypertension. These techniques do not aim at specifically controlling blood pressure but rather aim at facilitating a general relaxation response. In two unpublished studies (Love, Montgomery, & Moeller, Note 1; Montgomery, Love, & Moeller, Note 2) subjects were trained in EMG feedback, progressive relaxation, and autogenic training; Patel (1973, 1975a) and Patel and North (1975) combined EMG and GSR feedback with various relaxation/meditation techniques.

Review of the Literature on the Psychological Control of Essential Hypertension

The 21 most recent and important single-group and between-groups studies in the field are reviewed and summarized in table form. They have been grouped into the following four categories: (a) blood pressure biofeedback studies (Table 1); (b) relaxation/meditation studies (Table 2); (c) mixed studies, that is, studies in which various feedback and relaxation/meditation techniques have been combined (Table 3); and (d) comparative studies, that is, studies comparing blood pressure biofeedback and relaxation/meditation training (Table 4). These tabular presentations are followed by a detailed methodological critique including suggestions for future research and a discussion of results.

Methodological Critique

A careful inspection of the tables clearly reveals that most of the reviewed studies

have methodological faults, many of them serious ones. A definitive evaluation of the benefits of psychological procedures in the control of essential hypertension is therefore not possible. The following methodological discussion not only points out the faults in these studies but also suggests methodological requirements for future research.

Subject selection. In many of the studies, subject-selection criteria were not reported. For example, length of time a person had been diagnosed as essential hypertensive, age limits, and the required blood pressure level for inclusion into the experiment were often not defined. Some researchers worked with young subjects in whom hypertension is a rare phenomenon (e.g., the average age of subjects in the Stone and DeLeo, 1976, study was 28 years) or with borderline hypertensives (e.g., Shoemaker & Tasto, 1975; Surwit, Shapiro, & Good, 1978). Also, in several cases the diagnosis of essential hypertension was not verified (e.g., Elder & Eustis, 1975). It is recommended that subject samples be homogeneous with respect to (a) an established diagnosis of essential hypertension, (b) a minimum hypertension history, and (c) a minimum pretrial blood pressure.

Concurrent pharmacological treatment. Hypotensive medication and tranquilizers were being taken by subjects in 14 out of the 21 studies. In most studies, drug dosages were systematically stabilized or decreased during the training period. In some investigations, however, drugs or drug dosages were changed for medical reasons unrelated to the research project during the training period (Taylor, Farquhar, Nelson, & Agras, 1977) or during follow-up (Blackwell et al., 1976). This, of course, makes a meaningful interpretation of data difficult. It is possible that certain hypotensive drugs interact with various training procedures, but this has not yet been researched. Deabler, Fidel, Dillenkoffer, and Elder (1973) trained both a drug and no-drug group in progressive relaxation and hypnosis and found no apparent Drug × Training interaction. Unfortunately, no statistical analyses were performed, and the overall quality of the study was poor. One other aspect worth mentioning here is that subjects on medication who participate in biofeedback or relaxation/meditation training may change their drug-taking behavior as a result of the training. So far, drug compliance has not been investigated or reported in any of the relaxation/meditation or biofeedback studies. To avoid serious interpretative problems, it is recommended that investigators treat subjects who receive medication as a distinct group from those who do not.

Baseline assessment of blood pressure. As mentioned before, blood pressure responds to a great variety of environmental and internal factors and exhibits considerable diurnal and beat-to-beat variability. Dollery (1973) has shown that systolic blood pressure during a 24-hour period can be as low as 65 mmHg during sleep and as high as 170 mmHg (and over) during maximal exertion. Shapiro and Surwit (1976) mention that a series of measurements taken every half minute can have a range of up to 30 mmHg. It is well documented (Dunne, 1969; Pickering, 1968) that persons react to the taking of blood pressure with elevations in pressure. In the course of repeated assessments, as subjects adapt to the laboratory situation, blood pressure typically decreases.

In the studies reviewed, several researchers report no baseline period other than the measurement of blood pressure immediately prior to the first training session (Deabler et al., 1973; Elder & Eustis, 1975; Goldman, Kleinman, Snow, Bidus, & Korol, 1975; Kristt & Engel, 1975; Patel, 1973; Walsh et al., 1977; Hager & Surwit, Note 3), or they reported baseline measures for less than 1 week (Elder et al., 1973; Patel & North, 1975; Shoemaker & Tasto, 1975; Surwit et al., 1978). It is highly likely that in some studies, training effects were confounded with blood pressure variability and the effects of adaptation and resting. In single-case experimental research, measurements of blood pressure should be repeated until they stabilize. No clear guidelines have yet been established for group studies. However, it is safe to say that four assessment sessions over a 4-week period is a minimum requirement.

Control groups. The arguments given

previously not only suggest the use of extended baseline assessment of blood pressure but also the inclusion of a waiting list control group. Type, length, and number of assessment sessions should be identical for all groups throughout the baseline and training or control period. A further reason for the inclusion of a control group is that the mere fact of being attended to, the fact that something is being done about their prob-

Table 1
Psychological Control of Essential Hypertension: Biofeedback Studies

n	Concurrent pharmacological treatment	Baseline	Training procedure
		Benson et al. (1971)	
7	6 subjects on stable dosage of hypotensive medication	Average of 11 sessions over 2 weeks (sessions run until BP stabilized)	Systolic BP feedback: Constant cuff method, contingent slide projection, and monetary rewards
		Elder et al. (1973)	
18	Several subjects on stable (?) dosage of various CNS depressants	1 session of 20 BP determinations	1. Diastolic BP feedback ($n = 6$): Visual signal every 2 min. contingent on diastolic BP reduction 2. Diastolic BP feedback, contingent verbal approval ($n = 6$) 3. Control group ($n = 6$): Asked to relax and lower BP, no feedback
		Elder & Eustis (1975)	
22	20 subjects on stable (?) dosage of psychotropic or hypotensive medication	No baseline period, pretrial baseline only	Diastolic BP feedback: Visual feedback (green light for changes below, red light for no change or increase above basal pressure) at 1-min. intervals, verbal reinforcement
		Kristt & Engel (1975)	
5	All subjects on stable dosage of hypotensive medication	No baseline period, pretrial baseline only	Systolic BP feedback: Constant cuff method, display of cumulative performance scores; Week 1: Raise systolic BP; Week 2: Lower systolic BP; Week 3: Alternately lower, raise, lower BP within single session

lem, may bring considerable relief to subjects, with subsequent falls in blood pressure over time. Only 10 out of the 21 studies reported here included a control condition, but in 3 of these (Goldman et al., 1975; Patel, 1975a; Stone & DeLeo, 1976) subjects were not allocated in random fashion. In addition, several of the control conditions were inadequate. In some studies, especially those investigating within-session changes,

Table 1 (*continued*)

Frequency and duration of training (and follow-up)	BP changes pre/posttest and follow-up	Statistical analysis	Control
	Benson et al. (1971)		
Average of 22 45-min. sesssions on consecutive days (sessions run until no reduction for 5 sessions)	Within-session decrease: −17 mmHg systolic, 5 out of 7 subjects responded	Significant within-session reduction (ANOVA)	Extensive baseline; no control groups; no across-sessions measures; no follow-up
	Elder et al. (1973)		
7 40-min. sessions over 3 days, 1-week follow-up	1. 7% reduction in diastolic BP by Sessions 7 and 8 2. 20% reduction in diastolic BP in Sessions 3–8 3. No significant changes in BP	Group 2 significantly superior to Groups 1 and 3 in sessions 3–8; Group 1 superior to 3 in sessions 8 and 9 (ANOVA); reductions maintained at follow-up, (n = 11) no statistical tests	Control group; subjects started on salt-free diet 3 days prior to training; no extended baseline and follow-up; selective reporting and lack of detail in results section
	Elder & Eustis (1975)		
Spaced sessions (n = 19): 8 sessions of 20 trials each over 7 weeks, 1-month follow-up Massed sessions (n = 4): 10 sessions over 12 days, no follow-up	% difference from basal pressure presented in graphical form only Within-session reduction: Approximately 5% systolic + diastolic (somewhat larger for massed sessions) Follow-up: Within-session reductions approximately 3% systolic and diastolic	Significant difference between first and second half of training sessions (Mann-Whitney U)	No properly verified diagnosis of essential hypertension; no control groups; data based on % difference from basal pressure, which was assessed only once in first training session; data reported in confusing way
	Kristt & Engel (1975)		
42 sessions of 3 blocks of 10 trials each over 3 weeks, 3-month follow-up	Subjects reliably increased and decreased systolic BP, ability maintained at follow-up Pretest follow-up BP reductions as recorded by 4 subjects at home: −18.2 mmHg systolic −7.5 mmHg diastolic	Trend analysis statistics significant for all 3 conditions No statistical tests reported	No control groups; no clinic assessment of pre-follow-up changes in BP reported

(*table continued on pages 1024–1025*)

Table 1 (*continued*)

n	Concurrent pharmacological treatment	Baseline	Training procedure
		Goldman et al. (1975)	
11	None	No baseline period, pretrial baseline only	1. Systolic BP feedback (n = 7): Constant cuff method 2. Control group (n = 4): Asked to relax in their own best way
		Kleinman et al. (1977)	
8	None	3 2-hour clinic sessions over 3 weeks, BP recording by subjects 5 times daily over 2 weeks	Systolic BP feedback: Constant cuff method

Note. BP = blood pressure; CNS = central nervous system; ANOVA = analysis of variance. Reported

control subjects either did not sit for the same length of time as subjects in the training conditions (Deabler et al., 1973; Shoemaker & Tasto, 1975) or attended fewer sessions than the training group (Goldman et al., 1975; Taylor et al., 1977).

As in all forms of psychological intervention, nonspecific treatment factors undoubtedly play an important role in biofeedback and relaxation/meditation training. Various experiments (e.g., Goldring, Chasis, Schreiner, & Smith, 1956; Grenfell, Briggs, & Holland, 1964) have suggested that placebo treatments can have dramatic effects on blood pressure, although these results were confounded by the fact that treatment coincided with the beginning of longer hospitalization periods. In the present review only two studies included controls for nonspecific treatment effects. Taylor et al. (1977) used nondirective discussion groups in which subjects monitored and explored life's tensions and discussed solutions. Frankel, Patel, Horowitz, Friedwald, & Gaardner (1978) compared noncontingent diastolic blood pressure feedback with contingent feedback and also used a waiting list control group. In both studies no significant blood pressure reductions were observed in the control groups. Unfortunately, it was not determined whether the nonspecific treatment control conditions were as credible as the actual training procedures. As Kazdin and Wilcoxon (1976) pointed out, subjects who are exposed to control conditions that are less credible than the treatment condition are less likely to expect improvement. Expectancy of improvement is one of the most powerful nonspecific treatment factors, and it can be hypothesized that the apparent beneficial

Table 1 (*continued*)

Frequency and duration of training (and follow-up)	BP changes pre/posttest and follow-up	Statistical analysis	Control
	Goldman et al. (1975)		
1. 9 2-hour sessions of 30 trials each over 9 weeks 2. 3 2-hour sessions over 3 weeks	1. Within-sessions: −7 mmHg systolic Across-sessions: −6 mmHg systolic −15 mmHg diastolic 2. Within-session: −1 mmHg systolic Across-sessions: 4 mmHg systolic −4 mmHg diastolic	Within-session: Reductions significant for Group 1 only (ANOVA) Across-sessions: Changes significant for Group 1 on diastolic BP only (ANOVA)	Groups 1 and 2 had different pretraining BP, unsophisticated BP measurement; no random allocation; different number of sessions given to Groups 1 and 2; no follow-up
	Kleinman et al. (1977)		
9 2-hour sessions of 25–30 trials each over 9 weeks; follow-up BP recorded by subjects up to 4 months	Within-session: a. Control session: −.5 mmHg systolic −1.6 mmHg diastolic b. Feedback session: −4 mmHg systolic −4 mmHg diastolic Across-sessions (as recorded by subjects): −8 mmHg systolic −9 mmHg diastolic Maintenance of BP reduction at follow-up (n = 3)	Within-session: Reduction significant for Group b only on systolic BP (t test) Across-sessions: Reductions significant for systolic and diastolic BP (t test) No statistical test on follow-up data	No control groups; no report of laboratory-recorded across-sessions changes

blood pressure changes refer to across-session changes unless otherwise specified.

effect of a particular treatment may simply be due to this effect (Borkovec & Nau, 1972). For future research it is therefore important not only to include nonspecific control groups but also to ascertain that there is equal expectancy of improvement across conditions (Steinmark & Borkovec, 1974).

Length of training and follow-up period. A further drawback of many of the studies is that they use short training periods of sometimes not more than 1 week (Deabler et al., 1973; Elder et al., 1973) or use only a small number of training sessions. In studies reporting negative or statistically significant but clinically irrelevant results, it is therefore difficult to evaluate whether the technique was inefficient or the training was not intensive enough. In addition, follow-up assessment was frequently omitted or covered too short a period of time. The long-term durability of training effects is clearly a crucial issue that needs much more careful attention in future research. A minimum training period of 3 months and a follow-up period of 1 year is therefore suggested. Even longer follow-ups are necessary to determine whether and how the psychological control of essential hypertension effects morbidity and mortality.

Home practice and life-style changes. Sixteen out of 21 researchers asked their subjects to practice their respective techniques daily at home. Often, however, details of home practice procedures were not given, and only one researcher (Frankel et al., 1978) reported on actual compliance rates. Despite this lack of systematic data, there is evidence which suggests that regular daily home practice is crucial in achieving

and maintaining reductions in blood pressure. The commitment involved is similar to if not greater than that in pharmacological treatment, in which medication is usually taken on a regular daily basis and is a lifelong commitment. It is recommended that home

Table 2
Psychological Control of Essential Hypertension: Relaxation/Meditation Studies

n	Concurrent pharmacological treatment	Baseline	Training procedure
		Deabler et al. (1973)	
21	9 subjects on stable dosage of hypotensive medication assigned to drug group	No baseline period, pretrial baseline only	1. Progressive relaxation and hypnosis ($n = 6$) 2. Progressive relaxation, hypnosis, and drugs ($n = 9$) 3. Control group ($n = 6$) 7 BP checks over 4–5 day period
		Benson et al. (1974a)	
22	None	About 6 sessions over 6 weeks	TM
		Benson et al (1974b)	
14	All subjects on stable dosage of hypotensive medication	About 6 sessions over 6 weeks	TM
		Blackwell et al. (1976)	
7	All subjects on stable dosage of hypotensive medication throughout training period; dosage changes during follow-up	Up to 10 sessions over period of up to 10 weeks	TM
		Stone & DeLeo (1976)	
19	None	14 determinations over 10–14 days	1. Breath-counting meditation ($n = 14$) 2. Control group ($n = 5$) 6 BP checks over 6-month period

practice frequencies be recorded and reported as data in all future research.

It is worth adding that subjects who regularly practice a relaxation/meditation technique may as a consequence undertake changes in their life-style. To date, such

Table 2 (*continued*)

Frequency and duration of training (and follow-up)	BP changes pre/posttest and follow-up	Statistical analysis	Control
		Deabler et al. (1973)	
8–9 sessions over 4–5 days	Within session reductions in last session: 1. −17 mmHg systolic −19 mmHg diastolic 2. −16 mmHg systolic −14 mmHg diastolic 3. no significant changes	Within-session reductions significant for Groups 1 and 2 (ANOVA); no statistical tests perfomed to compare groups	Selective reporting of results; no across-sessions comparisons; training period too short; controls received pretrial readings only and did not sit for same length of time; no follow-up
		Benson et al. (1974a)	
6 sessions of 1–1½ hours each (and checking sessions) over 6 months	−7 mmHg systolic −4 mmHg diastolic	Reductions significant for systolic and diastolic BP (*t* test)	Subjects were self-selected; no control groups; no follow-up
		Benson et al. (1974b)	
6 sessions of 1½–2 hours each (and checking sessions) over 5 months	−11 mmHg systolic −5 mmHg diastolic	Reductions significant for systolic and diastolic BP (*t* test)	Subjects were self-selected; no control groups; lack of detail in results section; no follow-up
		Blackwell et al. (1976)	
6 sessions of 1½–2 hours each (and 10 checking sessions) over 9–12 weeks, 6-month follow-up	Clinic measures: −4 mmHg systolic −2 mmHg diastolic Follow-up: −3 mmHg systolic −4 mmHg diastolic	No statistical tests performed for the whole group	Recording of BP at home and clinic; no control groups; changes in drug treatment during follow-up
		Stone & DeLeo (1976)	
5 20-min. sessions over 6 months	1. Supine: −9 mmHg systolic −8 mmHg diastolic Upright: −15 mmHg systolic −10 mmHg diastolic 2. Supine: 1 mmHg systolic 2 mmHg diastolic Upright: −2 mmHg systolic ±0 mmHg diastolic	Reduction significant for Group 1 only on mean arterial pressure (*t* test)	Assessment of various biochemical variables; independent observer; no random allocation; no follow-up

(*table continued on pages 1028–1029*)

Table 2 (*continued*)

n	Concurrent pharmacological treatment	Baseline	Training procedure
		Taylor et al. (1977)	
31	All subjects on hypotensive medication, changes during training period for 14 subjects	2 sessions over 2 months	1. Medication control group (n = 11) 2. Nonspecific therapy (n = 10): Nondirective discussion groups (and self-monitoring) 3. Progressive relaxation (n = 10), breathing, imagery exercises and self-monitoring
		Pollack et al. (1977)[a]	
20	9 subjects on stable dosage of hypotensive medication	No. of sessions not reported, 3-month baseline period	TM

Note. BP = blood pressure; TM = transcendental meditation; ANOVA = analysis of variance.

potential changes (e.g., diet, exercise, smoking, alcohol consumption, use of drugs) have not been monitored but may well make some contribution to the reduction of blood pressure.

Assessment and reporting of training effects. The quality of studies was highly variable both in terms of their assessment procedures and the completeness of reports. In this context it is important to point out the difference between within-session and across-sessions measures. Within-session measures are typically taken either while the person is practicing a self-regulatory technique or at the end of the training session when blood pressure is likely to be lowest. To determine whether results achieved under training conditions have generalized to nontraining conditions, blood pressure has to be measured and reported across sessions, that is, independent of or immediately prior to the self-regulatory practice. All blood pressure values presented in Tables 1–4 were based on across-sessions measures unless otherwise specified.

In several studies, baseline blood pressure values were compared with those at the end of the last training session (Elder & Eustis, 1975; Elder et al., 1973; Goldman et al., 1975; Patel, 1973; Walsh et al., 1977). This of course, creates biased results that are likely to reflect both the effects of adaptation and resting and those of specific training.

In other studies, across-sessions measures were simply not reported (Benson et al., 1971; Deabler et al., 1973; Elder et al., 1973; Elder & Eustis, 1975; Kleinman et al., 1977), although these data must have been taken at least in the biofeedback studies, if only for calibration purposes. Finally, some experimenters relied on blood pressure values as recorded by the subjects themselves without reporting any assessment of relia-

Table 2 (*continued*)

Frequency and duration of training (and follow-up)	BP changes pre/posttest and follow-up	Statistical analysis	Control
	Taylor et al. (1977)		
For Groups 2 and 3, 5 30-min. sessions over 8 weeks, 6-month follow-up	1. −1 mmHg systolic ±0 mmHg diastolic Follow-up: −7 mmHg systolic −2 mmHg diastolic 2. −3 mmHg systolic −2 mmHg diastolic Follow-up: −4 mmHg systolic −4 mmHg diastolic 3. −14 mmHg systolic −5 mmHg diastolic Follow-up: −12 mmHg systolic −6 mmHg diastolic	Group 3 superior to Groups 1 and 2 on systolic BP only, significant difference between Groups 1 and 2 No significant difference between groups for systolic and diastolic BP at follow-up (sign test)	Independent observers; therapist unaware of results; no waiting list control group; medication changes; unsophisticated measurement of BP
	Pollack et al. (1977)[a]		
6 sessions of 1½–2 hours each (and 10 checking sessions) over 6 months	After 3 months: −10 mmHg systolic −2 mmHg diastolic After 6 months: −6 mmHg systolic −2 mmHg diastolic	Only reduction of systolic BP after 3 months significant, all other comparisons not significant	Measurement of plasma-renin activity; no control groups; incomplete reporting; no follow-up

[a] All values are approximate.

bility or comparative laboratory results (Kleinman, Goldman, Snow & Korol, 1977; Kristt & Engel, 1975; Hager & Surwit, Note 3).

For future research the simultaneous use and detailed reporting of within-session and across-sessions measures are strongly recommended. Temporary blood pressure reductions that occur only during training are of limited clinical importance. On the other hand, exclusive reliance on across-sessions measures prohibits the acquisition of information on the short-term effectiveness of a given technique. The concomitant use of both measures is particularly relevant in understanding failure to respond to training. Regarding the quality of blood pressure recording, researchers are urged to employ independent observers and to use either automated recording devices (Krausman, 1975; Tursky, 1974) or the random-zero-sphygmomanometer (Wright & Dore, 1970).

Generalization of training effects. The generalization of training effects is an important issue that has received much discussion but has not yet been systematically studied. Although several experimenters used different settings for training and assessment (Benson, Rosner, Marzetta, & Klemchuck, 1974a, 1974b; Blackwell et al., 1976; Frankel et al., 1978; Pollack, Weber, Case, & Laragh, 1977; Taylor et al., 1977), in the majority of studies reviewed here, training effects were assessed in the same environment in which training took place. It can therefore be argued that training effects were specific to the particular experimenter and to the laboratory setting in which subjects were trained and that in some cases the obtained changes would not have been maintained in a nonlaboratory environment. So far this latter issue has not been scientifically investigated.

Further, all training and assessments have

Table 3
Psychological Control of Essential Hypertension: Mixed Studies

n	Concurrent pharmacological treatment	Baseline	Training procedure
		Patel (1973)	
20	All subjects on hypotensive medication, 12 subjects stopped or reduced medication during experiment	No controlled baseline assessment	1. Various relaxation and meditation techniques and GSR feedback (n = 20)
		Patel (1975a)	
20	All subjects on hypotensive medication, 2 subjects stopped medication during experiment	No controlled baseline assessment	2. Control group (medication only) (n = 20): ½ hour resting instead of training
		Patel & North (1975)	
34	All subjects on stable dosage of hypotensive medication	3 sessions on 3 separate days	1. Various relaxation and meditation techniques, GSR and EMG feedback, and self-control procedures (n = 17) 2. Control group (medication only) (n = 17): ½ hour resting instead of training
		Frankel et al. (1978)[a]	
22	7 subjects on stable dosage of hypotensive medication	8 determinations over 6–8 weeks	1. Diastolic BP and EMG feedback, autogenic training, and progressive relaxation (n = 7): Constant cuff method 2. Placebo treatment (n = 7): Noncontingent diastolic BP feedback only 3. Control group (n = 8): Weekly BP checks only

Note. BP = blood pressure; GSR = galvanic skin response; EMG = electromyograph.

been conducted in a sitting or recumbent position in an undemanding environment under conditions of drastically reduced external stimulation. Although blood pressure taken under standard resting conditions gives us useful information, this type of data tells us nothing about whether training effects have generalized to the person's everyday environment or not. Blood pressure changes, especially the speed and magnitude of elevations and their rate of recovery, are also of importance. It is hoped that this crucial issue of generalization of training effects will be approached soon. It requires the use of portable monitoring equipment, which is now available (Littler, Honour, Pugsley, & Sleight, 1975; Littler, Honour, Sleight, & Stott, 1972). One perhaps more

Table 3 (*continued*)

Frequency and duration of training (and follow-up)	BP changes pre/posttest and follow-up	Statistical analysis	Control
	Patel (1973)		
36 ½-hour sessions over 3 months	1. Reduction from first to last training session (BP taken during training session): −20 mmHg systolic −13 mmHg diastolic	Reduction significant for systolic + diastolic BP (*t* test)	No independent observer
	Patel (1975a)		
9-month follow-up, Group 2 same as Group 1, 9-month follow-up	1. Follow-up: −15 mmHg systolic −13 mmHg diastolic 2. −1 mmHg systolic −2 mmHg diastolic Follow-up: 1 mmHg systolic −1 mmHg diastolic	1. No statistical test on follow-up data 2. Reductions on systolic and diastolic BP not significant (*t* test)	No random allocation, medication changes during follow-up
	Patel & North (1975)		
12 ½-hour sessions over 6 weeks, 3-month follow-up for Group 1 only	1. −26 mmHg systolic −15 mmHg diastolic Reduction maintained at follow-up 2. −9 mmHg systolic −4 mmHg diastolic	Reductions significant for Groups 1 and 2 on systolic and diastolic BP; Group 1 superior to Group 2 (*t* test); no statistical tests performed on follow-up data	Independent observer; control group, baseline period too short
	Frankel et al. (1978)[a]		
20 sessions over 4 months	1. 3 mmHg systolic 1 mmHg diastolic 2. −1 mmHg systolic −2 mmHg diastolic 3. 5 mmHg systolic 1 mmHg diastolic	No significant change in BP across sessions (*t* test) No significant differences between groups No significant within-session reductions in Groups 1 or 2	BP assessment by independent observer in different locality outside laboratory; control groups; no follow-up

[a] For all measures, subjects were supine.

practical alternative is to expose subjects to standard stressors under laboratory conditions. To date only one investigator (Patel, 1975b) has studied the effects of biofeedback-aided relaxation/meditation training on blood pressure response and recovery. In a preliminary experiment, 32 essential hypertensives were randomly assigned to either a training or a waiting list control condition. Maximum blood pressure elevations in response to a "stressful" exercise and cold pressor test and the time taken for recovery were assessed before and after 6 weeks of training. Relaxation training resulted in significant reductions in systolic and diastolic blood pressure elevations and recovery time for both tests. Control subjects did not display any significant reductions on either

(*text continued on p. 1034*)

Table 4
Psychological Control of Essential Hypertension: Comparative Studies

n	Concurrent pharmacological treatment	Baseline	Training procedure
		Shoemaker & Tasto (1975)	
15	Not reported	3 sessions of 10 BP determinations each over 6 days	1. Progressive relaxation (n = 5) 2. Systolic and diastolic BP feedback (n = 5): Direct feedback from BP chart recorder at 90-sec intervals 3. Control group (n = 5): 6 BP checks over 2 weeks
		Surwit et al. (1978)	
24	50% on stable (?) dosage of hypotensive and psychotropic medication	2 1-hour sessions over 1 week	1. Combined BP and heart rate feedback (n = 8) Constant cuff method, feedback contingent on simultaneous reduction of BP and heart rate (20 1-min. trials) 2. Frontalis and extensor EMG feedback (integrated) (n = 8) 3. Breath-counting meditation (n = 8) and information of BP at end of session
		Hager & Surwit (Note 3)	
30	Unspecified no. of subjects on stable dosage of hypotensive medication	No baseline data reported	1. Systolic BP feedback (n = 15) Visual feedback and performance score counter (portable home practice unit) 2. Breath-counting meditation (n = 15)
		Walsh et al. (1977)	
24	12 subjects on stable (?) dosage of hypotensive and psychotropic medication	No baseline period, pretrial baseline only	Phase 1: (n = 24) 1. Progressive relaxation 2. Pulse wave velocity feedback Phase 2 (n = 16 of the 24 Phase 1 subjects) 3. Training 1 and 2 combined

Note. BP = blood pressure; EMG = electromyograph; ANOVA = analysis of variance.

Table 4 (*continued*)

Frequency and duration of training (and follow-up)	BP changes pre/posttest and follow-up	Statistical analysis	Control
	Shoemaker & Tasto (1975)		
6 80-min. sessions over 2 weeks	1. −7 mmHg systolic −8 mmHg diastolic 2. −1 mmHg systolic −1 mmHg diastolic 3. 2 mmHg systolic 1 mmHg diastolic	Significant within-session and across-sessions reductions in systolic and diastolic BP for Group 1 only (linear trend comparison)	Low pretraining BP values; no extended baseline and follow-up; control subjects did not sit for same length of time as Groups 1 and 2
	Surwit et al. (1978)		
8 sessions of 1–1½ hours each over 4 weeks, 6-week follow-up, 1-year follow-up	1. 5 mmHg systolic Follow-up: 1 mmHg systolic 2. 6 mmHg systolic Follow-up: −4 mmHg systolic 3. −6 mmHg systolic Follow-up: −8 mmHg systolic	No significant reduction within or across sessions; no significant between-groups differences (ANOVA) 1-year follow-up: Average BP for all groups combined: 3 mmHg systolic 3 mmHg diastolic (no statistical analysis performed)	Low pretraining BP values; careful matching procedures; baseline period too short; no control group
	Hager & Surwit (Note 3)		
20 min. twice daily home practice sessions over 4 weeks (self-record of BP before and after each session)	Within-session reduction for Groups 1 and 2 combined (n = 17): −4 mmHg systolic −2 mmHg diastolic Across-sessions reductions for Groups 1 and 2 combined (n = 17): −2 mmHg diastolic	No significant differences between Groups 1 and 2 on all comparisons (ANOVA) Within-session reductions significant on systolic and diastolic BP (sign-test) Across-sessions reductions significant for diastolic BP only (ANOVA)	No clinic data (i.e., exclusive reliance on BP as recorded by subjects); no control group; no baseline or follow-up, high dropout rate: 8 in Group 1, 5 in Group 2
	Walsh et al. (1977)		
5 1-hour sessions over 5 weeks, 5 sessions of 7 3-min. trials each, information on pre/post session BP Training period not reported, 5 1½-hour sessions, 3-month and 1-year follow-up	Phase 1 No details given; graphical description only Phase 2 Slight increases in BP (no details given) Groups 1 and 2 (combined) reduced BP from 146/94 mmHg at beginning of Phase 1 to 134/87 mmHg at end of Phase 2 3-month follow-up: Subjects from Group 1 had lower systolic BP 1-year follow-up: No between-groups difference	Within-session: Group 2 superior to Groups 1 on diastolic BP (ANOVA) Across-sessions: No significant between-groups difference (ANOVA) No statistical analysis reported Significant difference (t test) No statistical analysis reported	No control group; incomplete reporting; errors due to technical problem?

measure. With the exception of systolic blood pressure rises in the exercise test, differences were significant for all between-groups comparisons. Despite some methodological shortcomings, these results are highly relevant and worth replicating.

Discussion of Results

In the following section, the effectiveness of the various techniques that have been applied to the control of essential hypertension is critically evaluated. Each of the three approaches presented in Tables 1–3, namely, blood pressure biofeedback, relaxation/meditation, and biofeedback combined with relaxation/meditation, is discussed separately and highlighted by a brief description of the most noteworthy studies. Finally, blood pressure and relaxation/meditation training are compared (Table 4), and explanations for differences in outcome are proposed.

Blood pressure biofeedback. Ten studies investigating blood pressure biofeedback in the treatment of essential hypertension are presented in Tables 1 and 4. In general, results support findings of studies using normotensives as subjects. Under laboratory conditions various blood pressure feedback techniques produced significant across-sessions reductions in systolic blood pressure (ranging from 6–18 mmHg) and in diastolic blood pressure (ranging from 8–15 mmHg). Two studies stand out: Benson et al. (1971) and Kristt and Engel (1975).

Benson et al. (1971) trained seven subjects using the constant cuff technique. Baseline assessment and training sessions were individualized, and training continued until each subject showed no further reductions in systolic blood pressure in 5 consecutive sessions. The average number of training sessions was 22 (range = 8–34) over a 4½-week period, and average within-session decreases in systolic blood pressure were 17 mmHg. Although this study did not include a control group and lacks across-sessions and follow-up data, it is one of the best controlled in the field.

Kristt and Engel (1975) taught five subjects to reliably increase and decrease systolic blood pressure within sessions. Again the constant cuff technique was used. Data were presented in graphical form only but suggested average increases and decreases in systolic pressure of 10–15 mmHg. Training was conducted in 42 sessions during a 3-week hospital stay. After discharge from the hospital, subjects continued to practice lowering systolic blood pressure at home on a daily basis. They did this by inflating a standard cuff to their average blood pressures and making Korotkoff sounds disappear while cuff pressure was maintained. Cuff pressure was then adjusted to the new lower pressure, and the procedure was repeated. After 3 months of home training, blood pressure as measured by subjects in their home, had decreased 18 mmHg systolic and 8 mmHg diastolic from pretreatment baseline. For several reasons these results have to be interpreted cautiously: (a) Subjects had shown no across-sessions reductions in the laboratory; (b) blood pressure taken by subjects at home was not checked against blood pressure taken by an independent observer; (c) there was no control group, and the number of subjects at follow-up was too small (n = 4); and (d) there is no way to determine whether the simulated blood pressure control procedure was instrumental in producing these reductions or whether other factors, such as taking time out to practice and relax, were the crucial components.

Two studies (Shoemaker & Tasto, 1975; Surwit et al., 1978) failed to produce any decreases in blood pressure. In both cases the feedback techniques that were used appear inappropriate. Shoemaker and Tasto applied a noncontinuous proportional feedback technique in which once every 90 sec subjects were shown their systolic and diastolic blood pressure on a chart recorder and asked to reduce both. This probably was too complex a task. In the study by Surwit et al., which used the constant cuff technique, subjects received binary feedback for simultaneous reductions in blood pressure and heart rate; that is, feedback for a "correct" response was only given when decreases in blood pressure coincided with decreases in heart rate. As the authors themselves pointed out, this procedure was

probably ineffective because subjects had normal heart rates to start with.

With regard to which blood pressure feedback technique is most effective, no conclusive answer is possible. Comparative studies with essential hypertensives have not yet been conducted. But there is evidence which suggests that the constant cuff technique is the most promising one, whereas the pulse wave velocity technique still needs further testing. However, at present we have no convincing indication that essential hypertensives can achieve clinically relevant and persistent blood pressure reductions through blood pressure feedback training. Successful blood pressure control may require more intensive individualized training (e.g., Benson et al., 1971) and continued home practice (Kristt & Engel, 1975). Although home practice appears to be crucial for the maintenance of blood pressure reductions in relaxation/meditation training, the role of home practice and its relaxation components in blood pressure feedback training is far from clear (Tarler-Benlolo, 1978).

Relaxation/meditation. The overall methodological quality of relaxation/meditation studies is better than that of blood pressure biofeedback studies. Subject samples are larger, and across-sessions measures and adequate follow-up or extended training periods of approximately 6 months are reported in all but one study (Deabler et al., 1973). Blood pressure reductions across sessions were in the range of 7–14 mmHg for systolic and of 4–10 mmHg for diastolic blood pressure. Two studies are worth describing in some detail.

Stone and DeLeo (1976) compared 14 subjects who were trained in breath meditation with a small control group ($n = 5$). Training consisted only of five sessions, but subjects were asked to practice regularly at home. After 6 months of practice, systolic and diastolic blood pressure in the supine position were reduced by 15 and 10 mmHg, respectively, whereas blood pressure in the control group had slightly increased (+1/+2 mmHg). Plasma dopamine β-hydroxylase levels were also found to be significantly reduced and to correlate with falls in blood pressure. Dopamine β-hydroxylase is an enzyme that converts dopamine to norepinephrine and has been suggested to be an indicator of sympathetic nervous system activity. The fact that the experimental and the control group were of unequal size and that subjects were not randomly allocated to groups necessitates caution in interpretation. In addition, pretreatment blood pressures were in the borderline hypertensive range, and subjects were much younger than in all other studies reviewed here.

In an exceptionally well-controlled study, Taylor et al. (1977) compared three groups of subjects on medication ($n = 31$) over an 8-week period. The first group received relaxation training, the second, nonspecific treatment, and the third, no treatment at all. The relaxation technique consisted of progressive muscle relaxation, the imagination of pleasant scenes, and self-monitoring. The relaxation group achieved the greatest decreases (−14 mmHg systolic, −5 mmHg diastolic), but only in the case of systolic blood pressure were they statistically significant. At a 3-month follow-up assessment, relaxation subjects had generally maintained their reductions (−12/−6 mmHg) but did not differ significantly from the other two groups, which had by then displayed further small decreases in blood pressure (−7/−2 mmHg for the medication-only and −4/−4 mmHg for the nonspecific treatment group).

Several single-case studies, which so far have not been mentioned in this review, have also investigated the effects of progressive relaxation and the modification thereof on essential hypertension (Beiman, Graham, & Ciminero, 1978; Bloom & Cantrell, 1978; Brady, Luborsky, & Kron, 1974; Graham, Beiman, & Ciminero, 1977). The report by Brady et al. (1974) deserves special mention because it used A-B-A and A-B-A-B single-case designs that have otherwise not been used in clinical studies of essential hypertension. (For a discussion of single-case designs in clinical biofeedback research, see Barlow, Blanchard, Hayes, & Epstein, 1977.) After 2–4 weeks of daily half-hour blood pressure assessment sessions (Baseline 1), the three subjects in the Brady et al. study received between 19 and 25 half-hour training sessions of metronome-conditioned relaxation training

(Training 1) over 4 weeks. This was followed by another 4 weeks of daily blood pressure checks without any further training practice (Baseline 2). One subject showed no change, whereas the other two showed significant decreases in diastolic blood pressure (of 3 mmHg and 6 mmHg, respectively) during Training 1, and significant increases in diastolic blood pressure (of 5 mmHg and 8 mmHg, respectively) in Baseline 2. One subject resumed relaxation after Baseline 2, which resulted in a drop in diastolic blood pressure of 13 mmHg. Systolic blood pressure was not reported.

Different relaxation and meditation techniques have not yet been compared with each other; therefore, at this stage, little can be said about which technique is likely to be more effective. Of the three techniques that have been investigated, progressive relaxation and breadth meditation appear to have produced slightly larger decreases in blood pressure than TM. Of the seven studies reviewed in Table 2, none of the four TM studies included a control group.

Mixed studies. Four studies that investigated a combination of various biofeedback and relaxation/meditation techniques in the psychological control of essential hypertension are reviewed in Table 3. The work by Patel and North (1975) has shown the most impressive results of all studies discussed in this review. Training consisted of a combination of educational programs, rhythmic slow breathing, muscular relaxation, meditation, GSR and EMG feedback, and various self-control procedures for coping with everyday difficult situations. Reductions in mean values after a 6-week training and a 3-month follow-up period were 26 mmHg systolic and 15 mmHg diastolic. The 17 control subjects who attended the same number of sessions but simply rested and relaxed on their own also achieved significant reductions of 9 mmHg systolic and 4 mmHg diastolic. However, the difference between the two groups was highly significant. The results are particularly convincing in that the investigators used a half-crossover design in which control subjects later underwent the full training procedure and as a result reduced their blood pressure by 28 mmHg systolic and 16 mmHg diastolic.

Frankel et al. (1978) used a similar training package consisting of diastolic blood pressure feedback (constant cuff technique), EMG feedback, autogenic training, and progressive relaxation and compared it with a group receiving noncontingent diastolic blood pressure feedback and with a waiting list control group. After 4 months of training, no significant blood pressure reductions were found in any of the groups. The reasons for these diametrically opposed outcomes are difficult to assess. Because of the simultaneous application of several training procedures, it is impossible to isolate the active ingredients in these training packages. However, training appeared to be more intensive and comprehensive in the Patel study, and baseline systolic blood pressure for the whole sample was also considerably higher (168/100 vs. 153/99).

Blood pressure biofeedback and relaxation/meditation compared. The four studies comparing blood pressure biofeedback and relaxation/meditation techniques (Table 4) have not helped much in assessing the possible differential effectiveness of these techniques. Two of the studies slightly favor progressive relaxation over noncontinuous, proportional (Shoemaker & Tasto, 1975), and pulse wave velocity feedback (Walsh et al., 1977). The other two studies both find breath meditation equally ineffective as a portable constant cuff technique (Hager & Surwit, Note 3) and as a constant cuff technique for simultaneous reductions in systolic blood pressure and heart rate (Surwit et al., 1978). Unfortunately, the studies of Walsh et al. and Hager and Surwit are of poor methodological quality, and the studies by Shoemaker and Tasto and Surwit et al. probably used, as previously discussed, inappropriate blood pressure feedback techniques. Finally, a recent experiment with normotensives (Steptoe, 1978), which compared pulse wave velocity feedback and breath meditation, also produced equivocal results.

In drawing general conclusions regarding the comparative effectiveness of relaxation/meditation and blood pressure feedback techniques, we are therefore restricted to the non-

comparative studies previously reviewed. Although experimental evidence is still somewhat weak, it is safe to say that in contrast to blood pressure biofeedback training, relaxation/meditation training has produced small but significant reductions in blood pressure with essential hypertensives. These reductions have been shown to persist for up to 6 months and to generalize to environments other than those in which training was conducted (e.g., Frankel et al., 1978). To account for the greater effectiveness of relaxation/meditation approaches, the following explanations are suggested.

1. Integrated physiological pattern versus specificity of blood pressure control. More effective, rapid, and persistent blood pressure control is likely to be attained if the self-regulation of blood pressure is accompanied by compatible changes in other physiological parameters such as heart rate, respiratory rate, muscle relaxation, and so on. Relaxation/meditation techniques are assumed to be more effective because they elicit such an integrated physiological pattern (Schwartz, 1976), whereas blood pressure feedback training may only affect blood pressure without concomitant changes in heart rate and other physiological parameters.

2. Home practice. As a rule relaxation/meditation techniques are taught in five to eight sessions, but the major part of training actually takes place in the subject's home environment in which he or she is advised to practice daily. These home practice sessions, which are usually not part of blood pressure biofeedback training, may be an important factor in explaining the superiority of relaxation/meditation approaches. It is interesting to note that the only blood pressure feedback study that showed blood pressure reductions to persist at a 3-month follow-up was also one of the few studies in which subjects practiced reducing blood pressure at home on a daily basis (Kristt & Engel, 1975).

3. Patient involvement. The role of patient involvement in the psychological control of essential hypertension has so far not been systematically studied. But an unpublished report by Sherman and Gaardner (Note 4) suggests that this factor may affect treatment outcome. They rated all available studies using biofeedback and relaxation/meditation techniques according to their treatment effectiveness and the degree of patient involvement. They found a significant correlation indicating that the more patients were involved, the higher was the treatment effectiveness. Patient involvement was defined by the number of training sessions, the intensity of home practice and recording, awareness of unwanted stress, expectancy of improvement elicited by training instructions, and the number of techniques used. Although Sherman and Gaardner did not differentiate between blood pressure feedback and relaxation/meditation training, it appears that the latter would receive higher ratings on patient involvement than the former.

Several other differences between blood pressure feedback and relaxation/meditation training with regard to patient involvement are worth mentioning here. In relaxation/meditation training subjects are taught individually or in groups by an instructor who is likely to respond to individual questions, who gives encouragement, and who may serve as a model. In blood pressure biofeedback training, the patient–instructor contact is probably much shorter and less personal than in relaxation/meditation training. It is mainly the "machine" that is in the role of the instructor. Training is also likely to take place in a more technical, controlled, and alien environment. In addition, repeated cuff inflations and the maintenance of cuff pressure may be unpleasant and distracting for some subjects.

Finally, relaxation/meditation approaches have the advantage of also positively affecting stress-related complaints such as insomnia (e.g., Borkovec & Hennings, 1978) and other clinical problems (Shapiro & Giber, 1978) and of increasing self-reported measures of health, performance, and well-being (Peters, Benson, & Porter, 1977). Subjects experiencing such changes may well become highly motivated to continue home practice on a regular basis.

If techniques of relaxation/meditation are combined with self-control procedures and with EMG and GSR feedback training, blood pressure effects are likely to be even larger (Patel & North, 1975). The combined use of

diastolic blood pressure feedback training and other techniques for relaxation has produced negative results (Frankel et al., 1978). On the other hand, Fey and Lindholm (1978), who worked with normotensive subjects, found that blood pressure feedback (constant cuff technique), when combined with progressive relaxation training, produced larger within-session reductions in systolic blood pressure than progressive relaxation alone. However, in a follow-up session, when no feedback was provided, the two groups no longer differed. Feedback training then appears to improve or facilitate within-session control of blood pressure but tends to lose its effect when it is withdrawn.

Issues of Experimental Design

All of the studies that have been reviewed in Tables 1–4 are outcome studies of either between-groups or single-group designs, except for one study of single-case experiments (Benson et al., 1971). The inadequacies of the single-group design are obvious and need no further mention here. Accordingly, the discussion concentrates on the pros and cons of between-groups versus single-case experimental designs. Two major advantages of the between-groups comparison design stand out. (a) It allows the investigation of the comparative effectiveness of two or more training procedures, and (b) results have a wider application. In contrast, generalizations based on single-case experiments have to be made cautiously.

Large individual differences in response to relaxation/meditation and biofeedback training have been reported by many researchers. Single-case experiments have the great advantage of treating intrasubject and intersubject variability not as error but as information (Barlow et al., 1977). One consequence of intersubject variability is of particular concern in group designs, as Hersen and Barlow (1976) have pointed out:

> When broad divergence . . . occurs among clients in response to an intervention, statistical treatments will average out the clinical effects along with changes due to unwanted sources of variability. In fact, this type of intersubject variability is the rule rather than the exception. (pp. 37–38)

Group designs in general are based on the assumption that all subjects in a given sample have homogeneous characteristics. In the case of clinical samples, homogeneity of disorder and etiology is also assumed. However, since it is generally accepted that essential hypertension is a multicausal disorder, it is likely that within each group there are subsamples that respond differently to training. One way of overcoming this difficulty is to apply more stringent sampling criteria. Another alternative is to match subjects on certain variables. But unfortunately, at present, the relevant matching variables are simply not known. Group designs also imply that a standardized method of intervention is appropriate for all subjects. But because of the heterogeneity of etiology in essential hypertension, such an assumption is clearly inappropriate. Yet training procedures used in group designs must be designed in such a way that they are generally suitable for all subjects, and once the experiment is under way, they cannot be adjusted to the individual needs of the subject. It is then not surprising that many group outcome studies yield statistically significant but clinically irrelevant results. In contrast, single-case experiments allow for training procedures to be tailored to the specific needs of each person and permit the testing of specific components of a given training procedure.

For future research the use of both types of designs is recommended. Single-case experiments would prove particularly valuable in deciding which training procedure is likely to achieve the best results with which type of subject. This could be followed up by multifactorial between-groups studies in which one or more training procedures would be compared with different subject samples (e.g., comparing the effects of progressive relaxation and breath meditation on essential hypertensives with high versus low pretest frontalis EMG levels). Further, between-groups studies can be improved by not only gathering pretest, posttest, and follow-up data but also by repeating measurements during the training period. Such a repeated measures design (Kiesler, 1971) would allow both outcome and process information.

Suggestions for a Comprehensive Approach to Assessment and Training

It will have become clear from this review that so far, psychological approaches have had only limited success in controlling essential hypertension. One reason for this may be that the commonly invoked psychophysiological model of essential hypertension is too narrow and that assessment and treatment based on this model are too simplistic. Most studies attempting to control elevated blood pressure through psychological techniques have aimed at counteracting or reducing the sympathetic nervous system activity that is assumed to play a central role in essential hypertension. The question of what actually causes sympathetic arousal in the first place is rarely asked. It is hypothesized here that to a considerable degree, the person's sympathetic arousal reflects idiosyncratic ways of perceiving, appraising, and interacting with his or her environment. It could therefore prove useful not only to measure blood pressure and other related physiological variables but also to measure cognitive appraisal and coping patterns in response to environmental demand. Assessment could take place in the laboratory applying standard stressors (Richter-Heinrich, Knust, Müller, Schmidt, & Sprung, 1975) or in the person's real-life environment (e.g., at the person's work place). The latter is a practicable proposition because the necessary technology is now available (Littler et al., 1975).

Such comprehensive assessment would allow the determination of which situations and events the person responds to with blood pressure elevations and whether they are mediated by dysfunctional patterns of thinking, emoting, and behaving. A more comprehensive approach to training would make use of specific behavioral techniques, cognitive restructuring (Goldfried, 1977), and stress management techniques such as anxiety management (Bloom & Cantrell, 1978) and stress inoculation training (Meichenbaum, 1977). It is hoped that future research will examine whether these techniques that aim at systematically teaching active coping skills produce greater blood pressure effects than the more passive techniques of blood pressure control described in this review.

Conclusion

In conclusion, psychological approaches to the self-regulation of high blood pressure are a promising *adjunct* to pharmacological treatment, and under the supervision of a physician, may allow the gradual reduction of medication requirements as the patient becomes more proficient in his or her respective technique (Patel & North, 1975). However, without more large-scale clinical trials with sound methodology, an unqualified acceptance of these techniques as an *alternative* to pharmacological treatment is not justified. To date, few studies have resulted in blood pressure reductions that are clinically relevant by virtue of either their magnitude or duration. It has been clearly shown that substantial blood pressure reductions can be achieved under laboratory conditions and that after periods of consistent relaxation/meditation practice, these changes are maintained under resting conditions without prior practice. But it has not yet been demonstrated whether training reduces blood pressure in the person's real-life environment. The active therapeutic ingredients of the various techniques have not yet been satisfactorily established, and only two studies have used nonspecific treatment control procedures.

It is unlikely that any one single technique will be suitable for all subjects, but rather a variety of approaches is necessary to deal effectively with individual differences in physiological and psychological responding. It is important that research should concentrate on differential diagnosis, with the aim of establishing criteria for choosing the most appropriate intervention techniques. Such research would also help to stimulate the development of new techniques or to suggest new combinations of existing ones.

A wide application of psychological approaches to the prevention and control of essential hypertension is unlikely in the immediate future. On the one hand, much more needs to be known about which tech-

nique or combination of techniques works best with whom. On the other hand, a wide application of these techniques requires a major shift in attitude of the general population and of the medical and health professions. The majority of people may prefer the easier course of medication rather than the more demanding process of training and daily practice. This attitude may not so much reflect a lack of willingness to take responsibility for one's own health as a lack of awareness of how personal habits effect it and how they may be changed.

It is hoped that psychologists will play an increasingly important role in heightening general awareness and in researching and teaching the skills conducive to health and the prevention of disease.

Reference Notes

1. Love, W. A., Jr., Montgomery, D. D., & Moeller, T. A. *Working paper number 1.* Ft. Lauderdale, Fla.: Nova University, Behavioral Sciences Center, 1973.
2. Montgomery, D. D., Love, W. A., Jr., & Moeller, T. A. *Working paper number 2.* Ft. Lauderdale, Fla.: Nova University, Behavioral Sciences Center, March 1974.
3. Hager, J. L., & Surwit, R. S. *Hypertension self-control with a portable feedback unit for relaxation.* Paper presented at the meeting of the Society for Psychophysiological Research, San Diego, October 1976.
4. Sherman, R. A., & Gaardner, K. R. *Patient involvement and treatment effectiveness in behavioral treatments of hypertension.* Paper presented at the meeting of the Biofeedback Society of America, Orlando, Fla., March 1977.

References

Barlow, D. H., Blanchard, E. B., Hayes, S. C., & Epstein, L. H. Single-case designs and clinical biofeedback experimentation. *Biofeedback and Self-Regulation,* 1977, *2,* 221–239.

Beiman, I., Graham, L. E., & Ciminero, A. R. Self-control progressive relaxation training as an alternative nonpharmacological treatment for essential hypertension: Therapeutic effects in the natural environment. *Behaviour Research and Therapy,* 1978, *16,* 371–375.

Benson, H. *The relaxation response.* New York: Morrow, 1975.

Benson, H., Rosner, B. A., Marzetta, B. R., & Klemchuck, H. M. Decreased blood pressure in borderline hypertensive subjects who practice meditation. *Journal of Chronic Diseases,* 1974, *27,* 163–169. (a)

Benson, H., Rosner, B. A., Marzetta, B. R., & Klemchuck, H. M. Decreased blood pressure in pharmacologically treated hypertensive patients who regularly elicited the relaxation response. *Lancet,* 1974, *1,* 289–291. (b)

Benson, H., Shapiro, D., Tursky, B., & Schwartz, G. E. Decreased systolic blood pressure through operant conditioning techniques in patients with essential hypertension. *Science,* 1971, *173,* 740–742.

Blackwell, B., et al. Transcendental meditation in hypertension: Individual response patterns. *Lancet,* 1976, *1,* 223–226.

Blanchard, E. B., & Young, L. D. Self-control of cardiac functioning: A promise as yet unfulfilled. *Psychological Bulletin,* 1973, *79,* 145–163.

Bloom, L. J., & Cantrell, B. Anxiety management training for essential hypertension in pregnancy. *Behavior Therapy,* 1978, *9,* 377–382.

Borkovec, T. D., & Hennings, B. L. The role of physiological attention focusing in the relaxation treatment of sleep disturbance, general tension, and specific stress reaction. *Behaviour Research and Therapy,* 1978, *16,* 7–19.

Borkovec, T. D., & Nau, S. D. Credibility of analogue therapy rationales. *Journal of Behavior Therapy and Experimental Psychiatry,* 1972, *3,* 257–260.

Brady, J. P. Metronome-conditioned relaxation: A new behavioral procedure. *British Journal of Psychiatry,* 1973, *122,* 729–730.

Brady, J. P., Luborsky, L., & Kron, R. E. Blood pressure reduction in patients with essential hypertension through metronome-conditioned relaxation: A preliminary report. *Behavior Therapy,* 1974, *5,* 203–209.

Brener, J. A. A general model of voluntary control applied to the phenomena of learned cardiovascular change. In P. A. Obrist, A. H. Black, J. Brener, & L. V. DiCara (Eds.), *Cardiovascular psychophysiology.* Chicago: Aldine, 1974.

Bulpitt, C. J., & Dollery, C. T. Side effects of hypotensive agents evaluated by a self-administrated questionnaire. *British Medical Journal,* 1973, *3,* 485–490.

Cannon, W. B. *The wisdom of the body.* New York: Norton, 1932.

Davidson, R. J., & Schwartz, G. E. The psychobiology of relaxation and related states: A multi-process theory. In D. I. Mostofsky (Ed.), *Behavior control and modification of physiological activity.* Englewood Cliffs, N.J.: Prentice-Hall, 1976.

Deabler, H. L., Fidel, E., Dillenkoffer, R. L., & Elder, S. T. The use of relaxation and hypnosis in lowering high blood pressure. *The American Journal of Clinical Hypnosis,* 1973, *16*(2), 75–83.

Dollery, C. T. Normal and raised arterial pressure: What is hypertension? In G. Onesti, K. E. Kim, & J. H. Moyer (Eds.), *Hypertension: Mech*

nisms and management. New York: Grune & Stratton, 1973.

Dunne, J. F. Variation of blood pressure in untreated hypertensive outpatients. *Lancet*, 1969, *1*, 391–392.

Elder, S. T., & Eustis, N. K. Instrumental blood pressure conditioning in outpatient hypertensives. *Behaviour Research and Therapy*, 1975, *13*, 185–188.

Elder, S. T., Longacre, A., Jr., Welsh, D. M., & McAfee, R. D. Apparatus and procedure for training subjects to control their blood pressure. *Psychophysiology*, 1977, *14*, 68–72.

Elder, S. T., Ruiz, Z. R., Deabler, H. L., & Dillenkoffer, R. L. Instrumental conditioning of diastolic blood pressure in essential hypertensive patients. *Journal of Applied Behavior Analysis*, 1973, *6*, 377–382.

Elder, S. T., Welsh, D. M., Longacre, A., Jr., & McAfee, R. D. Acquisition, discriminative stimulus control, and retention of increases/decreases in blood pressure of normotensive human subjects. *Journal of Applied Behavior Analysis*, 1977, *10*, 381–390.

Fey, S. G., & Lindholm, E. Systolic blood pressure and heart rate changes during three sessions involving biofeedback or no feedback. *Psychophysiology*, 1975, *12*, 513–519.

Fey, S. G., & Lindholm, E. Biofeedback and progressive relaxation: Effects on systolic and diastolic blood pressure and heart rate. *Psychophysiology*, 1978, *15*, 239–247.

Frankel, B. L., Patel, D. J., Horowitz, D., Friedwald, W. T., & Gaardner, K. R. Treatment of hypertension with biofeedback and relaxation techniques. *Psychosomatic Medicine*, 1978, *40*, 276–293.

Goldfried, M. R. The use of relaxation and cognitive relabeling as coping skills. In R. B. Stuart (Ed.), *Behavioral self-management*. New York: Bruner/Mazel, 1977.

Goldman, H., Kleinman, K., Snow, M., Bidus, D., & Korol, B. Relationship between essential hypertension and cognitive functioning: Effects of biofeedback. *Psychophysiology*, 1975, *12*, 569–573.

Goldring, W., Chasis, H., Schreiner, G. E., & Smith, H. W. Reassurance in the management of benign hypertensive disease. *Circulation*, 1956, *14*, 260–264.

Graham, L. E., Beiman, I., & Ciminero, A. R. The generality of the therapeutic effects of progressive relaxation training for essential hypertension. *Journal of Behavior Therapy and Experimental Psychiatry*, 1977, *8*, 161–164.

Grenfell, R. F., Briggs, A. H., & Holland, W. C. A double-blind evaluation of antihypertensive drugs. *Angiology*, 1964, *15*, 163–170.

Gribbin, B., Steptoe, A., & Sleight, P. Pulse wave velocity as a measure of blood pressure changes. *Psychophysiology*, 1976, *13*, 86–91.

Gutmann, M. C., & Benson, H. Interaction of environmental factors and systemic arterial blood pressure: A review. *Medicine*, 1971, *50*, 543–553.

Henry, J. P., & Cassel, J. C. Psychosocial factors in essential hypertension: Recent epidemiologic and animal experimental evidence. *American Journal of Epidemiology*, 1969, *90*, 171–200.

Hersen, M., & Barlow, D. H. *Single-case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press, 1976.

Jacobson, E. Variation of blood pressure with skeletal muscle tension and relaxation. *Annals of Internal Medicine*, 1939, *12*, 1194–1212.

Jacobson, E. *Modern treatment of tense patients*. Springfield, Ill.: Charles C Thomas, 1970.

Kannel, W. B. Recent highlights from the Framingham study. *Australian and New Zealand Journal of Medicine*, 1976, *6*, 373–386.

Kannel, W. B., & Dawber, T. R. Hypertensive cardiovascular disease: The Framingham study. In G. Onesti, K. E. Kim, & J. H. Moyer (Eds.), *Hypertension: Mechanisms and management*. New York: Grune & Stratton, 1973.

Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, *83*, 729–758.

Kiesler, D. J. Experimental designs in psychotherapy research. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An experimental analysis*. New York: Wiley, 1971.

Kleinman, K. M., Goldman, H., Snow, M. Y., & Korol, B. Relationship between essential hypertension and cognitive functioning II: Effects of biofeedback training generalize to nonlaboratory environment. *Psychophysiology*, 1977, *14*, 192–197.

Krausman, D. T. Methods and procedures for monitoring and recording blood pressure. *American Psychologist*, 1975, *30*, 285–294.

Kristt, D. A., & Engel, B. T. Learned control of blood pressure in patients with high blood pressure. *Circulation*, 1975, *51*, 370–378.

Littler, W. A., Honour, A. J., Pugsley, D., & Sleight, P. Continuous recording of direct arterial pressure in unrestricted patients. *Circulation*, 1975, *51*, 1101–1106.

Littler, W. A., Honour, A. J., Sleight, P., & Stott, F. D. Continuous recording of direct arterial pressure and electrocardiogram in unrestricted man. *British Medical Journal*, 1972, *3*, 76–78.

LoGerfo, J. P. Hypertension. Management in a prepaid health care project. *Journal of the American Medical Association*, 1975, *223*, 245–248.

Mahesh Yogi, M. *Transcendental meditation*. New York: Signet, 1968.

Meichenbaum, D. H. *Cognitive-behavior modification*. New York: Plenum Press, 1977.

Naranjo, C., & Ornstein, R. E. *On the psychology of meditation*. New York: Viking Press, 1971.

Onesti, G., Kim, K. E., & Moyer, J. H. *Hyperten-*

sion: Mechanisms and management. New York: Grune & Stratton, 1973.

Patel, C. H. Yoga and biofeedback in the management of hypertension. *Lancet*, 1973, *2*, 1053–1055.

Patel, C. H. 12-month follow-up of yoga and biofeedback in the management of hypertension. *Lancet*, 1975, *1*, 62–65. (a)

Patel, C. H. Yoga and biofeedback in the management of "stress" in hypertensive patients. *Clinical Science and Molecular Medicine*, 1975, *48*, 171–174. (Supplement) (b)

Patel, C. H., & North, W. R. S. Randomised controlled trial of yoga and biofeedback in management of hypertension. *Lancet*, 1975, *2*, 93–95.

Peters, R. K., Benson, H., & Porter, D. Daily relaxation response breaks in a working population: I. Effects on self-reported measures of health, performance and well-being. *American Journal of Public Health*, 1977, *67*, 946–953.

Pickering, G. W. *High blood pressure*. London: Churchill, 1968.

Pollack, A. D., Weber, M. A., Case, D. B., & Laragh, J. H. Limitations of transcendental meditation in the treatment of essential hypertension. *Lancet*, 1977, *1*, 71–73.

Redmond, D. P., Gaylor, M. S., McDonald, R. H., & Shapiro, A. P. Blood pressure and heart-rate response to verbal instruction and relaxation in hypertension. *Psychosomatic Medicine*, 1974, *36*, 285–297.

Richter-Heinrich, E., Knust, U., Müller, W., Schmidt, K. H., & Sprung, H. Psychophysiological investigations in essential hypertensives. *Journal of Psychosomatic Research*, 1975, *19*, 251–258.

Schwartz, G. E. Self-regulation of response patterning: Implications for psychophysiological research and therapy. *Biofeedback and Self-Regulation*, 1976, *1*, 7–30.

Shannon, B. J., Goldman, M. S., & Lee, R. M. Biofeedback training of blood pressure: A comparison of three feedback techniques. *Psychophysiology*, 1978, *15*, 53–59.

Shapiro, D. A monologue on biofeedback and psychophysiology. *Psychophysiology*, 1977, *14*, 213–227.

Shapiro, D., Schwartz, G. E., & Tursky, B. Control of diastolic blood pressure in man by feedback and reinforcement. *Psychophysiology*, 1972, *9*, 296–304.

Shapiro, D., & Surwit, R. S. Learned control of physiological function and disease. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, N.J.: Prentice Hall, 1976.

Shapiro, D., Tursky, B., Gershon, E., & Stern, M. Effects of feedback and reinforcement on the control of human systolic blood pressure. *Science*, 1969, *163*, 588–590.

Shapiro, D. H., Jr., & Giber, D. Meditation and psychotherapeutic effects. *Archives of General Psychiatry*, 1978, *35*, 294–302.

Shapiro, D. H., Jr., & Zifferblatt, S. M. Zen meditation and behavioral self-control: Similarities, differences, and clinical applications. *American Psychologist*, 1976, *31*, 519–532.

Shoemaker, J. E., & Tasto, D. L. The effects of muscle relaxation on blood pressure of essential hypertensives. *Behaviour Research and Therapy*, 1975, *13*, 29–43.

Smith, J. C. Psychotherapeutic effects of transcendental meditation with control for expectation of relief and daily sitting. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 630–637.

Stamler, J., Stamler, R., Riedlinger, W. F., Algera, G., & Roberts, R. H. Hypertension screening of 1 million Americans. *Journal of the American Medical Association*, 1976, *235*, 2299–2306.

Steinmark, S. W., & Borkovec, T. D. Active and placebo treatment effects on moderate insomnia under counterdemand and positive demand instructions. *Journal of Abnormal Psychology*, 1974, *83*, 157–163.

Steptoe, A. Blood pressure control: A comparison of feedback and instructions using pulse wave velocity measurements. *Psychophysiology*, 1976, *13*, 528–535.

Steptoe, A. Voluntary blood pressure reductions measured with pulse transit time: Training conditions and reactions to mental work. *Psychophysiology*, 1977, *14*, 492–498.

Steptoe, A. The regulation of blood pressure reactions to taxing conditions using pulse transit time feedback and relaxation. *Psychophysiology*, 1978, *15*, 429–438.

Steptoe, A., Smulyan, H., & Gribbin, B. Pulse wave velocity and blood pressure change: Calibration and application. *Psychophysiology*, 1976, *13*, 488–493.

Stone, R. A., & DeLeo, J. Psychotherapeutic control of hypertension. *The New England Journal of Medicine*, 1976, *294*, 80–84.

Stoyva, J. A psychophysiological model of stress disorders as a rationale for biofeedback training. In F. J. McGuigan (Ed.), *Tension control: Proceedings of the second meeting of the American Society for the Advancement of Tension Control*. Blacksburg, Va.: University Publications, 1976.

Stoyva, J., & Budzynski, T. Cultivated low arousal —An antistress response? In L. V. DiCara (Ed.), *Limbic and autonomic nervous system research*. New York: Plenum Press, 1974.

Surwit, R. S., Shapiro, D., & Good, M. I. Comparison of cardiovascular biofeedback, neuromuscular feedback, and meditation in the treatment of borderline hypertension. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 252–263.

Tarler-Benlolo, L. The role of relaxation in biofeedback training: A critical review of the literature. *Psychological Bulletin*, 1978, *85*, 727–755.

Taylor, C. B., Farquhar, J. W., Nelson, E., Agras, S. Relaxation therapy and high blood pressure. *Archives of General Psychiatry*, 1977, *34*, 339–342.

Travis, T. A., Kondo, C. Y., & Knott, J. R. Heart rate, muscle tension, and alpha production of transcendental meditators and relaxation controls. *Biofeedback and Self-Regulation*, 1976, *1*, 387–394.

Treichel, M., Clinch, N., & Cran, M. The metabolic effects of transcendental meditation. *The Physiologist*, 1973, *16*, 472.

Tursky, B. The indirect recording of human blood pressure. In P. A. Obrist, A. H. Black, J. Brener, & L. V. DiCara (Eds.), *Cardiovascular psychophysiology*. Chicago: Aldine, 1974.

Veterans Administration Cooperative Study Group on Antihypertensive Agents. Effects of treatment on morbidity in hypertension: I. Results in patients with diastolic blood pressure averaging 115 through 129 mmHg. *Journal of the American Medical Association*, 1967, *202*, 1028–1034.

Veterans Administration Cooperative Study Group on Antihypertensive Agents. Effects of treatment on morbidity in hypertension: II. Results in patients with diastolic blood pressure averaging 90 through 114 mmHg. *Journal of the American Medical Association*, 1970, *213*, 1143–1152.

Walrath, L. C., & Hamilton, D. W. Autonomic correlates of meditation and hypnosis. *The American Journal of Clinical Hypnosis*, 1975, *17*, 190–197.

Walsh, P., Dale, A., & Anderson, D. E. Comparison of biofeedback pulse wave velocity and progressive relaxation in essential hypertensives. *Perceptual and Motor Skills*, 1977, *44*, 839–843.

Woolfolk, R. L. Psychophysiological correlates of meditation. *Archives of General Psychiatry*, 1975, *32*, 1326–1333.

Wright, B. M., & Dore, C. F. A random-zero sphygmomanometer. *Lancet*, 1970, *1*, 337–338.

Received May 1, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1044–1049

Cognitive Behavior Modification: Misconceptions and Premature Evacuation

Michael J. Mahoney and Alan E. Kazdin
The Pennsylvania State University

Ledwidge's recent implication that cognitive behavior modification is "a step in the wrong direction" is examined and evaluated. Misconceptions about the alleged differences between cognitive and behavior therapists are noted, with particular emphasis on metaphysical dualism and dichotomous categorization. The classification of therapists according to the techniques that they employ is also questioned. It is argued that procedural overlap is common across schools of therapy and that categorical distinctions—when they are possible—are more likely to be valid when they are based on theoretical assumptions rather than on therapeutic techniques. The argument that behavior therapists rely almost exclusively on nonverbal means is challenged, as is the assertion that cognitively oriented therapists "rely chiefly on speech as the instrument of change." Finally, it is argued that Ledwidge's cautions about the continuing pursuit of cognitive-behavioral techniques are themselves premature and contrary to the commitment to empirical evaluation shared by both cognitive and less cognitive behavior therapists.

It is perhaps not surprising that the emergence of cognition in behavioral quarters has stimulated such strong and vituperative reactions. Within the last 2 years alone, there have been almost a dozen articles and papers attacking the "mentalistic" resurrection of cognitive processes in behavior modification (Goldiamond, 1976; Observer, 1977, 1978; Rachlin, 1977a, 1977b; Skinner, 1977; Wolpe, 1976a, 1976b, 1978). The sentiment of most of these writings is aptly summarized in a recent *Psychological Record* editorial (Observer, 1978):

> Cognitivism constitutes a counter-revolution to the behavioristic revoluton that promised to promote psychology to a scientific status . . . . (p. 157)

> Students of scientific psychology cannot but deplore the regressive tendencies of cognitive psychology. (p. 159)

In his recent evaluation of cognitive behavior modification, Ledwidge (1978) sounds a similar alarm regarding the apparent trend toward cognitive theorization in behavior modification:

> A wholesale conversion to cognitive methods on insufficient evidence could rob behavior therapy of its distinctiveness and lead to the abandonment of the more traditional behavioral techniques, the success of which have afforded (behavior) therapy the reputation it enjoys today. (p. 354)

After a sporadic review of the available literature, Ledwidge draws a mixture of conclusions that seem only tangentially related to extant evidence. It is the purpose of this brief article to point out some of the misconceptions in Ledwidge's review and to isolate some of the pivotal issues that seem to be developing along the interface of cognitive psychology and behavior modification.

Dualism and Dichotomy

Two of the most persistent myths that surround the developing interface are the notions that (a) cognitivism necessitates mentalism and that (b) a therapist (or the

Requests for reprints should be sent to Michael J. Mahoney, Department of Psychology, The Pennsylvania State University, 417 Bruce V. Moore Building, University Park, Pennsylvania 16802.

orist) is either exclusively cognitive or behavioral (but never the twain shall meet). The first assumption is apparent in Skinner's (1974) recent responses to the growing interest in cognitive psychology. He uses the terms "cognitive psychologist" and "mentalist" interchangeably:

> By attempting to move human behavior into a world of nonphysical dimensions, mentalistic or cognitive psychologists have cast the basic issues in insoluble forms. They have also probably cost us much useful evidence . . . . (p. 118)

This same sentiment is reiterated by Ledwidge (1978) when he states that "the present controversy is, of course, just another round in the centuries old mind–body debate" (p. 360). What is interesting is that the metaphysical issue of a nonphysical mind does not appear to differentiate groups that have been traditionally labeled as cognitive and behavioristic. In a recent survey involving 42 of the most eminent living contributors to behavior therapy and cognitive behavior modification, no significant differences in belief in the existence of a "mind" emerged (Mahoney, in press). There were also no differences in the extent to which these two groups emphasized the importance of experimental rigor in theory evaluation.

The second myth is more subtly defended by Ledwidge's (1978) insistence on the differentiation of cognitive and behavioral therapies—a distinction that he defends on the grounds that (a) cognitive-behavioral techniques do not attempt to directly modify behaviors and that (b) should these new techniques fail to improve on the therapeutic power of behavior therapy, they will "unjustifiably detract from the good reputation that behavior modification enjoys today" (p. 372). Interestingly, Ledwidge concedes that all forms of therapy involve some form of cognition:

> All forms of therapy, including medical treatment, are cognitive to the extent that the therapist must convince the client to cooperate in the suggested procedure before therapy can begin. In this trivial sense all . . . therapies . . . are cognitive. (p. 357)

One might take issue with the assumption that compliance and trust are "trivial" issues, but the basic point is that clients are presumed to think. What is apparently overlooked is the equally salient observation that therapists behave. At some level of analysis, all forms of psychotherapy involve behaviors on the part of a therapist that are intended to produce changes in the ongoing experiences of a client. Thus, one might also argue that all therapies are simultaneously cognitive and behavioral.

A more compelling illustration of this point is offered by Bandura's (1977a) timely emphasis on the distinction between procedures and processes. According to Bandura—and many other persons labeled "cognitive behavior modifiers"—the processes that govern human adjustment (and maladjustment) are cognitive in nature. (i.e., They involve attentional processes, aspects of information storage and retrieval, etc.) However, in almost comic irony, it now appears that behavioral procedures may be among the most powerful methods for activating those cognitive processes. Thus, if any clear distinction can be drawn, the major difference between cognitive and less cognitive behavior modifiers does not lie in their therapeutic procedures so much as in their rationale and selection of a given procedure in an individual case. The more cognitively oriented therapist is inclined to employ a behavioral procedure appropriate to the "cognitive restructuring" presumed to be required.

Misconceptions About Cognitive Therapy

This confusion regarding process and procedure is most apparent in discussions about the modification of verbal behavior. In an attempt to lump all nonbehavioral approaches into a broad category of psychotherapy, Ledwidge (1978) appears to lament the recognition of thoughts as therapeutic targets in behavior therapy:

> Slowly, but surely, over the years, behavior theory has become more cognitive. Cognitions, relabeled self-statements, are classed as behaviors. Whereas changing a person's mind was, and still is, considered psychotherapy, changing a person's self-statements passes for behavior therapy. (p. 354)

It is, of course, a special form of therapy, and Ledwidge prefers to call it cognitive therapy rather than cognitive behavior modification. The label is perhaps less important than the accurate assumption that "cognitive change (is) the active ingredient in treatment" (p. 356). Much less accurate, however, is the assertion that

> whereas behavior therapists attempt to change behavior directly by using mainly nonverbal means, cognitive therapists . . . rely chiefly on speech as the instrument of change. (p. 356)

This artificial distinction is further elaborated by Ledwidge's later reference to cognitive therapies as "verbal therapies." The error of such a dichotomy is discernible on at least two counts. On one hand, the behavior therapist is reliant on verbal communication during treatment—a point that is in direct variance with Ledwidge's "nonverbal" attribution. More to the point, however, is the fact that persons labeled cognitive behavior modifiers are explicit in their emphasis on behavioral performance as a primary means of challenging maladaptive beliefs. Cognitive theorists ranging from George Kelly to Albert Bandura have been almost uniformly consistent in their reliance on active motoric performance in therapy (cf. Bandura, 1977a, 1977b; Beck, 1976; Kelly, 1955; Mahoney, 1977b; Meichenbaum, 1977).

In several places, Ledwidge (1978) recognizes the problems of distinguishing cognitive behavior modification and behavior therapy and the "somewhat arbitrary decision of choosing where on a cognitive-behavioral dimension to place a cutoff between the two types of techniques" (p. 357). The dependence of many cognitive techniques on overt behavioral and nonverbal means of effecting therapeutic change would make this distinction almost impossible. The ambiguity of the distinction between cognitive and behavior therapy is not the fault of Ledwidge. The problem is in trying to make distinctions at the level of techniques. At this level, what different therapists actually do often overlaps considerably, a fact that traditionally has served as an impetus for eclecticism or integration of seemingly incompatible theoretical positions. We, too, would have difficulty in making decisions using a single continuum based on a cognitive-behavioral dimension that would divide techniques, although extremes might be identified with agreement. At the conceptual level, distinguishing different forms of therapy is a more straightforward task, since the theoretical allegiance of a particular technique is readily identified.

Perhaps for clinical psychology as a whole, the most important distinctions among therapies are not made at the technique or conceptual levels. A dimension slighted in Ledwidge's (1978) review that is the most significant in distinguishing therapies is the commitment to empirical research as the crucible for treatment evaluation. Cognitive behavior modification is firmly committed to the tenets and practices of contemporary behavioral research. These include careful specification of treatment ingredients, multiple operationism of outcome, recognition of overt behavior as a major measure of treatment efficacy, and so on. In this regard, distinctions between select techniques, whether they are called behavioral or cognitive, tend to diminish. In the final analysis, it will, or should, be the theoretically sound and empirically established techniques that are embraced by the field. What these are called, how these develop, and the purity of their philosophical heritage will be interesting but not of ultimate importance.

Empirical Status

In his discussion, Ledwidge (1978) concludes that the data supporting cognitive-behavioral approaches are "meager" in comparison with the "enormous body of research validating the effectiveness of behavior therapy procedures" (p. 370). One could, of course, question whether behavior therapy procedures have been so overwhelmingly validated (e.g., Kazdin & Wilcoxon, 1976; Kazdin & Wilson, 1978). Likewise, one could question the classification of such procedures as imagery and modeling as behaviorally (rather than cognitively) oriented in

tervention strategies. More pertinent, however, may be Ledwidge's assertion that "the more cognitive the technique, the less effective it is" (p. 370). This is in ironic contrast to his concluding remark that judgment on the relative effectiveness of cognitive-behavioral strategies must be deferred until adequate research is available. Both explicitly and implicitly, it is clear that Ledwidge has already developed the hunch that cognitive-behavioral techniques are "a step in the wrong direction." This verdict is questionable on at least two counts.

First, there are now over a dozen studies that suggest that cognitive parameters may enhance either the predictive validity or the therapeutic power of previous techniques (Bandura, 1977a; Mahoney & Arnkoff, 1978; Meichenbaum, 1977). The potential of cognitive perspectives is most apparent in the recent studies by Taylor and Marshall (1977) and Rush, Beck, Kovacs, and Hollon (1977). In the former it was shown that a cognitive-behavioral therapy was more effective than isolated cognitive, behavioral, or no-treatment groups in the management of mild to moderately depressed subjects. The Rush et al. study found that cognitive therapy was more effective than chemotherapy (tricyclic drugs) in the treatment of severe depression—a finding that is noteworthy in its constituting the first (and only) psychological treatment to surpass tricyclic drug efficacy in this realm. Likewise, it appears that cognitively dictated "mastery experiences" may enhance the effectiveness of behavior rehearsal and participant modeling techniques (Bandura, 1977a, 1977b).

A second reason for questioning Ledwidge's (1978) verdict is its prematurity. Cognitive-behavioral approaches are relatively recent developments, and adequate opportunity to evaluate their mettle has yet to offer itself. Just as Ledwidge has expressed concern about the "wholesale conversion to cognitive methods on insufficient evidence" (p. 354), one might question the wisdom of premature rejection of such methods on equally insufficient grounds. There are, of course, cognitive theorists who have been generous in their evaluation and optimism regarding the current status of cognitive-behavioral approaches (e.g., Ellis, 1977). One might find equally enthusiastic proponents of almost any system of psychotherapy. On the other side of the coin, however, it is worth acknowledging that many defenders of the developing cognitive-behavioral interface seem to be well aware of the challenges that continue to face this area of inquiry (e.g., Bandura, 1977b; Beck, 1976; Mahoney, 1974, 1977a, 1977b; Meichenbaum, 1977). At least some of these sentiments were expressed in the lead article of that new journal "devoted entirely to cognitive therapy" (Ledwidge, 1978, p. 359). After a conservative statement on the status of cognitive-behavioral strategies, the need for continuing self-scrutiny was emphasized:

> In sum, our long and arduous journey has just begun. Let us not waste time congratulating ourselves on our wisdom when our ignorance is still so salient. We have cause for optimism, but hardly for jubilation. There are throngs of suffering humans who fill our waiting rooms, and we have yet to demonstrate that our therapeutic promises will serve them better than have those of the past. Let us bear in mind that our ultimate commitment is to these persons, and not to the esoteric needs of our paradigm. We must be ready to change when so doing would serve them better, and we must be every ready to follow new paths toward clinical effectiveness. (Mahoney, 1977b, p. 15).

Conclusion

Among the many problems that arise in the therapy literature, two seem to be particularly objectionable and dangerous. The first and certainly foremost is overzealously advocating treatment techniques based on anecdotal information and weak case material. Therapy techniques continue to be hailed as effective in professional and lay circles as if empirical evidence attesting to their efficacy were already available. Scientific criteria for endorsing existing treatments have yet to be uniformly embraced in the psychotherapy literature, not to mention clinical practice.

The second problem is judging, in advance of empirical research, the kind of techniques that might be effective and ruling out certain avenues based on this judgment. Ledwidge's

(1978) article discusses the beginnings of research on the many different forms of cognitive behavior modification but at the same time warns about the potentially undesirable consequences of embracing these techniques. Those who object to cognitive behavior modification for reasons that they believe to be conceptual or metaphysical should rejoice in the fact that this area is strongly committed to empirical research. If in fact cognitive based techniques add so little to existing techniques, this will be demonstrated rather soon, since research in cognitive behavior modification is proliferating so actively. Alternatively, if this research proves to be heuristically and clinically productive, as current evidence suggests, critics will be required to examine their own criteria for decision making.

The current state of cognitive behavior modification calls for accelerated research rather than second thoughts over further exploration. Objections based on the possible conceptual and methodological impurities that cognitive theory or therapy might introduce reflect a failure to appreciate the historical lineage and contemporary characteristics of behavior therapy. Behavior therapy is hardly free from sin in the sense of having many loose theoretical ends, techniques based on concepts that stretch (and use) the imagination, and assertions of efficacy that are poorly based (Kazdin, 1978). To imply that cognitive theory or techniques somehow tarnish all of this is difficult to maintain.

Cognitive therapy has thrown itself into the evaluative arena of empirical research. This is a risk that many forms of traditional therapy have yet to take. Along with the risk of demise should be the potential benefits of successes and empirical insights. At the very least, critics as well as advocates might be well advised to suspend judgment until the data accrue.

References

Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 1977, *84*, 191–215. (a)

Bandura, A. *Social learning theory*. Englewood Cliffs, N.J.: Prentice-Hall, 1977. (b)

Beck, A. T. *Cognitive therapy and the emotional disorders*. New York: International Universities Press, 1976.

Ellis, A. Rational-emotive theory: Research data that supports the clinical and personality hypotheses of RET and other modes of cognitive-behavior therapy. *The Counseling Psychologist*, 1977, *7*, 2–42.

Goldiamond, I. Self-reinforcement as an explanatory fiction. *Journal of Applied Behavior Analysis*, 1976, *9*, 509–514.

Kazdin, A. E. *History of behavior modification*. Baltimore, Md.: University Park Press, 1978.

Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, *83*, 729–758.

Kazdin, A. E., & Wilson, G. T. *Evaluation of behavior therapy*. Cambridge, Mass.: Ballinger, 1978.

Kelly, G. A. *The psychology of personal constructs*. New York: Norton, 1955.

Ledwidge, B. Cognitive behavior modification: A step in the wrong direction? *Psychological modification*, 1978, *85*, 353–375.

Mahoney, M. J. *Cognition and behavior modification*. Cambridge, Mass.: Ballinger, 1974.

Mahoney, M. J. Cognitive therapy and research: A question of questions. *Cognitive Therapy and Research*, 1977, *1*, 5–16. (a)

Mahoney, M. J. Reflections on the cognitive-learning trend in psychotherapy. *American Psychologist*, 1977, *32*, 5–13. (b)

Mahoney, M. J. Cognitive and non-cognitive views in behavior modification. In P. O. Sjoden & S. Bates (Eds.), *Trends in behavior therapy*. New York: Academic Press, in press.

Mahoney, M. J., & Arnkoff, D. B. Cognitive and self-control therapies. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (2nd ed.). New York: Wiley, 1978.

Meichenbaum, D. *Cognitive-behavior modification*. New York: Plenum Press, 1977.

Observer. Concerning cognitive reversionism in psychology. *Psychological Record*, 1977, *2*, 351–354.

Observer. The recycling of cognition in psychology. *Psychological Record*, 1978, *28*, 157–160.

Rachlin, H. Reinforcing and punishing thoughts. *Behavior Therapy*, 1977, *8*, 659–665. (a)

Rachlin, H. A review of M. J. Mahoney's *Cognition and behavior modification*. *Journal of Applied Behavior Analysis*, 1977, *10*, 369–374. (b)

Rush, A. J., Beck, A. T., Kovacs, M., & Hollon, S. Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Therapy and Research*, 1977, *1*, 17–37.

Skinner, B. F. *About behaviorism*. New York: Knopf, 1974.

Skinner, B. F. Why I am not a cognitive psychologist. *Behaviorism*, 1977, *5*(2), 1–10.

Taylor, F. G., & Marshall, W. L. Experimental anal-

ysis of a cognitive-behavioral therapy for depression. *Cognitive Therapy and Research,* 1977, *1,* 59–72.

Wolpe, J. Behavior therapy and its malcontents—I. Denial of its bases and psychodynamic fusionism. *Journal of Behavior Therapy and Experimental Psychiatry,* 1976, *7,* 1–6. (a)

Wolpe, J. Behavior therapy and its malcontents—II. Multimodal eclecticism, cognitive exclusivism and "exposure" empiricism. *Journal of Behavior Therapy and Experimental Psychiatry,* 1976, *7,* 109–116. (b)

Wolpe, J. Cognition and causation in human behavior and its therapy. *American Psychologist,* 1978, *33,* 437–446.

Received May 4, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1050–1053

Cognitive Behavior Modification or New Ways to Change Minds: Reply to Mahoney and Kazdin

Barry Ledwidge
Simon Fraser University, Burnaby, Canada, and
Riverview Hospital, Port Coquitlam, Canada

The "myths" that (a) cognitivism necessitates mentalism and that (b) therapists can be classified on the basis of technique are defended. Two studies cited by Mahoney and Kazdin as evidence of the potential of cognitive perspectives are found, on close examination, to raise more questions than they answer. Charges of prejudgment and premature evacuation of the field are disclaimed, but it is urged that the phrase *behavior modification* not be used to describe cognitive approaches, since failure to distinguish the two kinds of therapy invites a conceptual confusion of cognition with behavior that could have unfortunate theoretical and practical consequences.

If it is possible to be damned with faint praise, then Mahoney and Kazdin's (1979) critique surely canonizes my position (Ledwidge, 1978) with faint criticism. Lacking adequate data, they fail to face squarely the central issues raised and instead attempt to shift the focus from treatment effectiveness to theoretical issues, including the old and bitter mind–body polemic. After dismissing as "sporadic" my review of the literature (which included a critical review of *every* article on cognitive-behavior modification published in any of the four behavior therapy journals between 1963, when the first issue of *Behaviour Research and Therapy* appeared, and July 1976, when I began writing the article) and after characterizing my conclusions as "only tangentially related to extant evidence," Mahoney and Kazdin then devote merely two paragraphs of their article to the empirical status of cognitive behavior modification and fail to cite any of the studies allegedly missed in my "sporadic review." The balance of their article is devoted to questions that are more semantic than real and that, by definition, cannot be answered empirically. It is to these pseudoissues that I now reluctantly turn.

The author would like to thank John Damron and Brian Cooney for their advice on substantive issues and comments on earlier drafts of this article.

Requests for reprints should be sent to Barry Ledwidge, Department of Psychology, Riverview Hospital, 500 Lougheed Highway, Port Coquitlam, British Columbia, Canada V3C 1J0.

Myths

I am taken to task by the authors for propagating two "persistent myths" that surround the "interface" of cognitive therapy and behavior modification. In the first place, according to Mahoney and Kazdin, I, like Skinner (1978) mistakenly assume that cognitivism necessitates mentalism.

What is a cognition, however, if not a mental event? A cognition is a behavior (albeit private), a hypothetical construct inferred from behavior to account for behavior, or a mental event in the bad sense of the word, that is, a causal event that is not reducible without remainder to either environmental, behavioral, or physiological events. The radical behaviorists, including Skinner, allow that private events (e.g., minute muscle movements detectable by an electromyograph) are acceptable for study but only if the investigator is able to reliably determine when the phenomenon occurs (Biglan & Kass, 1978). Verbal reports about cognitions qualify as behaviors in this sense, but the cognitions themselves do not. If cognitive behavior modifiers

were attempting merely to change the client's verbal reports about cognitions, then the treatment could be characterized as behavioral, but it is clearly the thoughts, beliefs, and attitudes of the clients and not just their speech acts that is the focus of treatment. Although radical behaviorists eschew the use of constructs (and theory building in general), methodological behaviorists are not averse to using intervening variables (e.g., Hull's r_g) to account for relationships among observable, environmental, and organismic events. If in fact cognitive behavior modifiers used cognition simply as a construct to account for enduring patterns of clients' responses to consulting room procedures, then there would be no confusion as to the status of a cognition. Unlike Hull, however, who never attempted to manipulate r_gs, Mahoney, Meichenbaum, Ellis, and others convert the intervening variable to a subject matter. (Mahoney and Kazdin, 1979, agree that "cognitive change is the active ingredient in treatment" [p. 1046].) If a cognition, then, as the term is used by cognitive behavior modifiers, is neither a behavior nor an intervening variable, then mental event remains as the most apt descriptor. To the extent that cognitive approaches appeal to causal events that are not environmental, behavioral, or specifically physiological, they are mentalistic. Hence, although Mahoney and Kazdin appear to repudiate psychophysical dualism, they are in a sense, tacit or "closet" dualists.

As evidence against my contention that cognitivism necessitates mentalism, Mahoney and Kazdin (1979) point out that in a soon-to-be-published survey (Mahoney, in press) involving 42 of the most eminent living contributors to behavior therapy and cognitive behavior modification, there were no significant differences in belief in the existence of a "mind." It does not surprise me that therapists of different persuasions have been able to accommodate in their lexicon this battered term. The question is whether the behavior therapists and the cognitive behavior modifiers surveyed mean the same thing when they say that mind exists. English and English (1958) list five definitions of mind, one of which might be acceptable to some behaviorists, namely,

> the organized totality or system of all **mental** processes or **psychic** activities, usually of an individual organism. The emphasis is upon the relatedness of the phenomena. Mind in this sense does not commit the user to a metaphysical position about the nature of these processes. Hence, it may be used by those who define psychology in terms of **acts** or **behaviors**. (p. 323; boldface in original)

The second "myth" that I am found guilty of propagating is that therapists can be categorized on the basis of the techniques they use. Mahoney and Kazdin (1979) point out that

> according to Bandura—and many other persons labeled "cognitive behavior modifiers"—the processes that govern human adjustment (and maladjustment) are cognitive in nature. (i.e., They involve attentional processes, aspects of information storage and retrieval, etc.) However, in almost comic irony, it now appears that behavioral procedures may be among the most powerful methods for activating those cognitive processes. Thus, if any clear distinction can be drawn, the major difference between cognitive and less cognitive behavior modifiers does not lie in their therapeutic procedures so much as in their rationale and selection of a given procedure in an individual case. The more cognitively oriented therapist is inclined to employ a behavioral procedure appropriate to the "cognitive restructuring" presumed to be required. (p. 1045)

The logic here is questionable. First we are told that cognitive behavior modifiers believe that the processes governing behavior are cognitive in nature. Mahoney (1977) asserts that the first premise of the cognitive-learning perspective is that "the human organism responds primarily to cognitive representations of its environments rather than to those environments per se" (p. 7). Then we are told that "in almost comic irony" (comical to behavior modifiers, ironic to cognitive behavior modifiers) behavioral techniques turn out to be the most efficient methods of changing these hypothetical constructs (an implicit endorsement of behavior therapy). Faced with data which indicate that human organisms respond primarily to the environment rather than to cognitive representations of it, Mahoney and Kazdin (1979) resolve this embarrassing theoretical paradox by asserting that although cognitive therapists and behavior therapists often use the same techniques, the best way

to distinguish them is on the basis of the rationale each uses for selecting a technique.

Surely, technique is the only logical basis for classifying therapists. Of what difference is it to the client whether the therapist's intent, when administering a procedure, was to reinforce discriminative operants, change irrational thought patterns, or strengthen the ego? Are we to believe that when participant modeling, administered by a cognitive therapist, results in a decrement in avoidance responding, it does so for different reasons (cognitive changes) than when it is used successfully by a therapist who does not consider cognitive change the active ingredient in participant modeling? Surely all therapists, including cognitive behavior modifiers, should base their choice of technique on empirical evidence of its effects on behavior and not on any theory associated with the technique.

Empirical Status

Two studies published after I had completed my review of the literature are cited as evidence of the potential of cognitive perspectives. One of these, Rush, Beck, Kovacs, and Hollon (1977), found that cognitive therapy was more effective than imipramine in the treatment of depressed outpatients. The treatment procedure used by the cognitive therapists is not specified, but we are referred to 43 pages of a book by one of the authors (Beck, 1976, pp. 263–305) and are told that

> the cognitive therapist employs both verbal and behavioral techniques to help the patient learn to (a) recognize the connections between cognition, affect, and behavior, (b) monitor his negative thoughts, (c) examine the evidence for and against his distorted cognitions, and (d) substitute more reality-oriented interpretations for his distorted negative cognitions. (pp. 18–19)

One wonders exactly what went on in these cognitive therapy sessions. Cognitive behavior modification may be firmly committed to the tenets and practices of contemporary behavioral research, but cognitive therapy does not lend itself to empirical investigation as well as does behavior therapy because some parameters of cognitive therapy cannot be specified. How does one operationally define, for example, a cognitive technique "to help the patient recognize the connections between cognition, affect, and behavior?" One wonders, too, how much of the cognitive therapy in the Rush et al. study consisted of verbal techniques and how much consisted of behavioral techniques. Would a straight behavior modification approach (e.g., differential reinforcement of depressive and nondepressive behavior by the patient's spouse or another significant person) have been more effective? These questions seem particularly relevant in the light of the findings of the second study cited by Mahoney and Kazdin (1979). Taylor and Marshall (1977) found that cognitive therapy and "behavioral intervention" (speech—verbal nonassertiveness—was the only behavior treated) were equally effective in reducing depression but that a combination of the two verbal approaches was more effective than either technique used singly. A behavior therapist might reasonably ask whether behavioral treatment of a broader spectrum of depressive behaviors (including speech) would prove superior to cognitive therapy or the combination treatment of Taylor and Marshall.

Conclusion

In their discussion, Mahoney and Kazdin (1979) accuse me of "judging, in advance of empirical research, the kind of techniques that might be effective and ruling out certain avenues based on this judgment" (p. 1047). The charge of prejudgment is supported by an out-of-context quote from my article (Ledwidge, 1978), in which I am alleged to have concluded that "the more cognitive the technique, the less effective it is" (p. 370). In context, the statement above turns out to apply only to one of three types of data reviewed (comparisons of two variants of the same behavior therapy technique in which one of the procedures relies less on cognitive operations than does the other and/or has extra behavioral components as part of the procedure). Seven such comparative studies were presented (including two by Bandura and his colleagues [Bandura, Blanchard, & Ritter, 1969; Bandura & Menlove, 1968]),

six of which showed that the more cognitive the technique, the less effective it was. Bandura (1977) has come to the same conclusion: "Regardless of the methods involved, results of comparative studies attest to the superiority of performance-based treatments" (p. 196).

At no point in my article did I recommend, as Mahoney and Kazdin (1979) imply, that research on the effectiveness of cognitive therapies be discontinued; in fact, more research and deferred judgment were called for:

> Existing CBM procedures may in time be shown to be as effective as behavior therapy with clinical disorders, or new and more effective cognitive methods may yet be devised. As documented earlier, cognitive behavior modification is of recent origin, and only a handful of studies have appeared in the journals so far. Until comparisons of behavior therapy and CBM are carried out with clinical populations, however, judgment on their relative effectiveness must be deferred (Ledwidge, 1978, p. 371)

The emphasis I intended to impart was that this new hybrid therapy should not be called cognitive behavior modification because it is not behavior modification. Cognitive therapists are engaged in a radical departure from the methodology of behaviorism in treating cognition as subject matter rather than as an intervening variable. I hoped to point out how failure to distinguish the two kinds of therapy invites a conceptual confusion of cognition with behavior that could have unfortunate theoretical as well as practical consequences.

If the requested name change sounds like school chauvinism, it is because it is. The hard-earned excellent reputation that behavior modification enjoys today would be tarnished if cognitive behavior modification proves no more effective than more traditional forms of psychotherapy.

References

Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review,* 1977, *84,* 191–215.

Bandura, A., Blanchard, E. B., & Ritter, R. The relative efficacy of desensitization and modeling approaches for inducing behavioral, affective, and attitudinal changes. *Journal of Personality and Social Psychology,* 1969, *13,* 173–199.

Bandura, A., & Menlove, F. L. Factors determining vicarious extinction of avoidance behavior through symbolic modeling. *Journal of Personality and Social Psychology,* 1968, *8,* 99–108.

Beck, A. T. *Cognitive therapy and the emotional disorders.* New York: International Universities Press, 1976.

Biglan, A., & Kass, D. J. The empirical nature of behavior therapies. *Behaviorism,* 1978, *5*(1), 1–15.

English, H. B., & English, A. C. *A comprehensive dictionary of psychological and psychoanalytical terms.* New York: McKay, 1958.

Ledwidge, B. Cognitive behavior modification: A step in the wrong direction? *Psychological Bulletin,* 1978, *85,* 353–375.

Mahoney, M. J. Cognitive therapy and research: A question of questions. *Cognitive Therapy and Research,* 1977, *1,* 5–16.

Mahoney, M. J. Cognitive and non-cognitive views in behavior modification. In P. O. Sjoden & S. Bates (Eds.), *Trends in behavior therapy.* New York: Academic Press, in press.

Mahoney, M. J., & Kazdin, A. E. Cognitive behavior modification: Misconceptions and premature evacuation. *Psychological Bulletin,* 1979, *86,* 1044–1049.

Rush, A. J., Beck, A. T., Kovacs, M., & Hollon, S. Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Therapy and Research,* 1977, *1,* 17–37.

Skinner, B. F. Why I am not a cognitive psychologist. *Behaviorism,* 1978, *5*(2), 1–10.

Taylor, F. G., & Marshall, W. L. Experimental analysis of a cognitive-behavioral therapy for depression. *Cognitive Therapy and Research,* 1977, *1,* 59–72.

Received September 21, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1054–1083

Psychobiology of Active and Inactive Memory

Donald J. Lewis
University of Southern California

A brief history and review of the short-term memory and long-term memory distinction is presented, and this distinction is concluded to be no longer adequate for either human or animal memory data. Simple memories for events are apparently formed quickly and are permanent. In such cases, an initial physiologically unstable period is not required. Thus, most forgetting is the result of a retrieval failure rather than a storage failure. A distinction between active memory (AM) and inactive memory (IM) is made. AM is a subset of IM and contains either newly formed memories or established retrieved memories or both. Some of the implications for psychobiology of the AM and IM distinction are discussed. It is suggested, for example, that while in AM, memories are particularly open to disruption either by amnesic agents or through other forms of interference. The forgetting process for new and established memories is time dependent (but independent of memory age) and is based on interference. It is desirable to maintain the distinction between memory storage and memory retrieval even while recognizing that associative storage aids in retrieval. The search for the biological basis of rapidly forming memories, perhaps based on the restructuring of protein fragments, remains important but the physiological brain processes underlying memory interference and retrieval require greater emphasis.

The distinction between short-term memory (STM) and long-term memory (LTM) has been a useful and productive one in both animal physiological psychology and human cognitive psychology in recent years. There is not total agreement on the characteristics of STM, but most agree that (a) STM is either the route of entry to LTM for new memories or at least the holding template until LTM processes are complete; (b) the contents of STM are temporary and fragile, either decaying rapidly, open to disruption, or transferring rapidly to LTM; and (c) the capacity of STM is limited to a few items at one time. LTM, on the other hand, stores memories relatively permanently and has almost unlimited capacity. There are many variations on these characteristics, some of which are discussed later. The purpose of this article is to show that the distinction between a temporary and fragile memory and one that can become permanent is no longer adequate for the preponderance of the data in either the animal or the human fields. A more adequate alternative is to consider all memories as long-term or permanent, with a few of them being active at any given time and the remainder of them in an inactive state. A brief history of the STM–LTM distinction will demonstrate why this is a more adequate alternative.

Precursors

William James (1890) introduced the term "primary memory," by which he meant the perceptions, images, sensory impressions, perceptual processes, concepts, and mental entities of any sort that occur as the result of external stimulations during learning. The items of primary memory are part of the psychological present. James used the term "specious present" to indicate that the con-

Requests for reprints should be sent to Donald J. Lewis, Department of Psychology, University of Southern California, University Park, Los Angeles, California 90007.

cept of the present was more apparent than real because what seemed like the present was overlaid by a great deal of the past and the future. This was part of James's firm insistence that consciousness was continuous and that any attempt to break it into parts was artificial. The "real" memory was his secondary memory, into which items could be stored and retrieved. For James, retrieved memories, even at the time of retrieval, were in secondary memory.

After James (1890), classical interest in primary memory focused on the number of items that could be simultaneously present. This line of research culminated in G. A. Miller's (1956) "magical number seven," consolidating the agreement among many studies that on the average, only seven items could be held in immediate memory and proposing the concept of "chunking," by which the amount of information could be greatly increased while retaining the limitation of approximately seven items.

Working in the field of verbal learning, Müller and Pilzecker (1900) attempted to explain the data on retroactive interference—that the learning of a second list of verbal materials causes the forgetting of a previously learned list. They argued for two physiological processes: first, a perseverating neural process that was maintained until, second, a more permanent memory structure was formed. During the perseverative phase, the neural process was open to disruption, and thus the second list of verbal materials disrupted the neural perseveration initiated by the first list before it could become a permanent physical structure. Müller and Pilzecker's "consolidation theory" has had an immense impact on neurological investigations of learning, even though its influence on verbal learning, for which it was originally proposed, has been minimal. A major reason for the neglect of perseveration theory by cognitive researchers has been the recognition of proactive interference; forgetting can be caused by the learning of verbal materials occurring long *before* the learning of materials whose forgetting ensues. Although retroactive interference could be due to a totally different forgetting mechanism than proactive interference, an unnecessary lack of parsimony would be required. Further, the finding that a major determinant of both retroactive and proactive interference was the similarity of the learned items turned researchers away from neurological concepts, since verbal similarity seemed difficult to conceptualize neurologically.

Ebbinghaus (1913) also proposed the possibility of two different memories, although his distinction was based on the number of items to be learned. If 6 or 7 nonsense syllables were to be committed to memory, only one trial was typically required, whereas 12 syllables required almost 15 repetitions. This discontinuity implied two different processes to Ebbinghaus, although it was not a distinction between an STM and an LTM, since both processes committed items to LTM.

Following Ebbinghaus (1913), researchers showed little interest in a STM–LTM distinction until the late 1940s, at which time both physiological psychologists and human cognitive psychologists became interested but with different orientations. The cognitivist interest grew out of research on information processing (Broadbent, 1958) and paid little attention to physiological work. Hebb (1949) renewed the interest of physiological researchers, who were in turn little concerned with the informational approach. In both fields there was an initial marked enthusiasm for two-process theories, an enthusiasm that later became tempered by skepticism and, in many cases, outright rejection.

Human Cognitive Conceptions

Experimental

Broadbent (1958) attached earphones to his human subjects and simultaneously presented two sets of three or four digits, one set to each ear, and the subjects were asked to recall each set. Broadbent found that subjects usually attended to only one set of digits and could recall this set perfectly. But they could also recall some of the second set even though the rate of presentation was too fast to permit the switching of attention from one ear to the other. From this, Broadbent concluded that there must be a temporary storage system, an STM, in which the second

set of digits was held momentarily, until attention could be turned to it. He conjectured that this storage system could hold the information only for a few seconds and that items that were not attended to within this time decayed and disappeared. Although Broadbent's STM mechanism was more complicated than has just been described, its essential features were (a) rapid decay rates, (b) continuing processing necessary to maintain items in STM, and (c) a limited capacity of only two to four items. Broadbent believed that these features of STM necessitated its distinction from LTM, which possessed none of these properties. Forgetting in LTM, he believed, was due to the interference produced by other memories and not to decay; items could be held in LTM indefinitely and without effort unless there was specific interference; the capacity of LTM was virtually unlimited.

Peterson and Peterson (1959) and Brown (1958) independently introduced both a new experimental procedure to study STM and additional data and theory to support the separation of STM from LTM. They presented a verbal unit consisting typically of three consonants—a trigram. If recall occurred immediately after a presentation, performance was usually perfect. If immediately after presentation of the trigram subjects were required to count backward by threes from a given number, there appeared a sharply decelerated "decay" function with retention almost at zero after an 18-sec delay. Here was good evidence for rapid forgetting, if the items could not be held continuously in STM, and Peterson and Peterson and Landauer (1974) supported a time-dependent decay theory of forgetting from STM, a notion very similar to that of the Müller and Pilzecker (1900) theory. This theory, you recall, had fallen into disuse in part because of the evidence that forgetting was produced by interference that was caused by the similarity between the target information and the interfering information. Further, this similarity could be defined on the bases of the semantic properties of the material. "House" and "dwelling," for example, would interfere greatly with each other, even though they had dissimilar physical features.

Thus, evidence presented by Conrad (1963) and Sperling (1963) that interference also occurred in STM was important. This interference, however, was acoustic in nature rather than semantic; for example, subjects would respond with an F when S was the correct trigram letter. Therefore, STM was still considered distinct from LTM, but acoustic interference replaced rapid decay in part, at least, as a distinguishing feature of STM. Important cognitive theories advanced at about this time incorporated conceptually different memory components, either as a physical store, a process, or both (e.g., Atkinson & Shiffrin, 1968; Glanzer, 1972; Waugh & Norman, 1965). Probably, by the end of the 1960s, a majority of the cognitive learning researchers accepted STM as real and distinct from LTM, even though the notion of a decaying or a consolidating engram (Müller & Pilzecker, 1900) never attracted a substantial number of adherents. (Nevertheless, see Hintzman, 1974, who considers consolidation a possible explanation for the "spacing effect," an improvement in retention when the interval between repetitions of certain verbal materials varies from 0 up to 15 sec.)

Amnesic Syndrome

Humans who have suffered damage to the hippocampus and/or areas around the walls of the third ventricle frequently show a profound learning and memory impairment of an extraordinary nature. Upon hearing a series of digits, they can repeat correctly as many as can normals, and they can carry on a normal conversation, which means they can hold in memory the early part of a message and use it to unitize the thought. They also appear to remember events that occurred before their injury. If they have a normal memory for current events and also for events prior to injury, why are they characterized as amnesics? They appear to be unable to transfer current memories to the long-term store. A conversation they have today, for example, will be forgotten tomorrow. STM and LTM each seem clear and distinct, but

new information does not appear to enter LTM. The most famous of these amnesics is H.M. (Milner, 1966, 1970). H.M. was a severe epileptic who underwent a bilateral hippocampectomi, after which he seemed unable to add any new information for later retrieval. No matter how many days in a row he had a conversation with a person, he could not remember the person's name or anything about the previous conversation on the next day. Other than this deficit, his behavior was not notably abnormal. There are many others who suffer from the amnesic syndrome as a result of brain damage, and, although their symptoms may vary slightly, they have in common an inability or difficulty in remembering events beyond a short period of time (Talland, 1968). This sort of dichotomized memory function seemed to demand a distinction between an STM and an LTM, and Atkinson and Schiffrin (1968), in their theoretical treatment of memory, characterized the amnesic syndrome as the single most convincing piece of evidence for a distinction between STM and LTM.

Arguments Against a Human STM–LTM Distinction

Experimental Evidence

Although the arguments and data supporting an STM–LTM distinction at the human level are strong, there has been considerable dissent from this position, a great deal of it centering around the Peterson-Brown (Brown, 1958; Peterson & Peterson, 1959) experimental procedure.

Melton (1963) had argued that one of the major distinguishing features of LTM was its openness to interference and that to the extent that a putative STM and the LTM showed identical interference effects, there would be no necessity to posit separate memories. Keppel and Underwood's (1962) demonstration that a considerable amount of the forgetting occurring in the Peterson-Brown procedure was due to proactive interference was thus a powerful support for a single-memory conceptualization. Then, Waugh and Norman (1965) (see also Glanzer, 1972; Wickelgren & Norman, 1966) showed that the number of interfering items was more important in producing forgetting in the short-term paradigm than the simple passage of time. The similarity among items has also been demonstrated as a cause of STM forgetting. Deutsch (1970), for example, produced greater forgetting of tone sequences when the interpolated material was also tonal than when words were used. Wickelgren, in a series of studies (e.g., 1965), showed both proactive and retroactive interference in STM when the learning and interfering materials were acoustically similar.

The evidence thus shows that both proactive and retroactive interference occur in STM and that this interference is the major cause of STM forgetting. Any simple decay (Reitman, 1974) that remains does not seem to be of much importance. Basically, these data seemed to support Melton's (1963) contention that if the putative STM obeys the same laws as LTM, there is no reason to assign it a separate existence. But interference may be acoustic in STM and semantic in LTM.

Assuming a separate STM, how is the transfer of memories made to LTM? Items are both maintained in STM and transferred to LTM through rehearsal. But rehearsal need not refer only to the holding through repetition of items in STM for a period of time. Craik and Watkins (1973) showed that simply maintaining items in STM for 10–20 sec did not improve their strength in LTM. (See Craik, 1979, for more detailed discussion.) Some form of coding is necessary to retain an item permanently. Ample evidence now exists that semantic coding does take place in STM (e.g., Shulman, 1970, 1972). There may still be a difference in the degree to which semantic coding takes place in STM and LTM, but the difference is not dichotomous and therefore neither are the memories on this basis. These coding studies have yielded a useful distinction between maintenance rehearsal and coding rehearsal. The simple holding or maintaining of an item in STM for a period of time, as has been indicated, aids little in its retrieval from LTM unless some coding operation is performed on it. It is the type of activity that occurs during

rehearsal that is crucial in STM itself, not the passage of time.

That different types of coding do occur in STM has helped to blur an STM–LTM distinction based on what has been called the "negative recency effect" (Madigan & McCabe, 1971). For immediate recall both the first and last items of several in a list are better retained than those in the middle. The usual explanation is that more attention and rehearsal can be devoted to the first items simply because they are first, whereas the last item is well retained because it is still in STM. Delayed recall, however, yields a negative recency because the last item received little rehearsal compared to the early items, and it is no longer in STM. But when appropriate coding is used (Bellezza & Walker, 1974), the usual positive recency is found also for delayed recall. Tulving (1968, 1974) has taken the view that there are two retrieval processes involved in recall rather than two storage processes. Early after item presentation, temporal and phonemic retrieval cues are effective. Later on, semantic cues are better. Murdock (1974) gives an intuitive example of this distinction. You are directing someone to pick up your suitcase from an airport conveyor discharge belt. "It is the one that has just come down the chute" would be one type of instruction. If it has been on the belt for some time, a different type of code would have to be used such as, "It is the blue one with the torn grip."

Based on the experimental data produced by cognitive learning psychologists, it seems fair to conclude that there is as yet no convincing reason to require a separate STM. But what of the evidence derived from the amnesic syndrome?

Amnesic Syndrome

H.M. had a normal immediate memory span, you recall, but he could not recognize people he had met many times or even recall yesterday's conversation. It was believed that he was unable to transfer memories from STM to LTM, supporting the notion of the independent existence of the two. Further studies, however, showed that H.M. could retain certain visual and tactile maze tasks (Corkin, 1968). H.M. also had some knowledge that he had a memory impairment (Milner, 1970), and he had learned a factory job, although he could not describe it. He knew that President Kennedy was assassinated, but he did not know who succeeded him. Perhaps one might conclude that H.M. has an aphasic deficiency in coding new information so that it can be recognized later. If so, the deficit is one of retrieval rather than of storage. Other evidence from amnesics lends support to this hypothesis in that with appropriate retrieval probes, memories can be recovered (Weiskrantz & Warrington, 1975). This point is discussed in greater detail later.

The phenomenon of retrograde amnesia (RA) is the memory loss caused by a brain trauma, a loss that is most severe for events just prior to trauma and is progressively less further into the past. This gradient is one of the bulwarks of consolidation theory (Glickman, 1961; Mah & Albert, 1973; McGaugh, 1966; Müller & Pilzecker, 1900), which requires new memories to be more open to disruption than older ones. As memories age, they harden and become relatively permanent. There are numerous observations, however, that are not consistent with the notion of a consolidating memory trace. (a) Many patients are amnesic for events that extend back a matter of years. (b) Shrinkage of amnesia is a frequent occurrence. The shrinkage indicates that during amnesia the memory existed but simply was not accessible. Shrinkage strongly suggests a failure to retrieve, not a failure to consolidate or to progress from STM to LTM. (c) RA frequently is spotted with islands of remembering that have no temporal continuity, and RA can involve condensing two events into one (Talland, 1968). The amnesia shrinks, but it is not a consistently temporal process moving forward in time. More islands appear, and only later does temporal coherence return. (d) Amnesics have difficulty remembering specific events but can remember well-categorized information or generalized information. They can, for example, remember that flags are for waving in a parade, but they do not remember a specific flag or a specific parade (Wood &

Kinsbourne, Note 1). (e) Amnesics frequently can learn motor skills and perceptual tasks even though they forget when or how they learned them (Milner, 1970; Warrington & Weiskrantz, 1970). (f) Proactive and retroactive interference can explain much of the forgetting of amnesics (Weiskrantz & Warrington, 1975).

In an excellent review of the amnesic syndrome, Kinsbourne and Wood (1975) argue that the reason that there appears to be a temporal gradient in many amnesics is that patients are asked different types of questions about the recent past than about the remote past. The questions about the remote past tend to be more general and categorical (semantic), except for the case history type of information that has undoubtedly been frequently rehearsed. Specific temporally related questions (episodic), on the other hand, are asked about recent events, and, since amnesics have trouble remembering specific events about any period of their life, there is the appearance of the forgetting of recent events. Kinsbourne and Wood related the forgetting that is characteristic of the amnesic syndrome to the distinction made by Tulving (1972) between episodic and semantic memory. Episodic memory is for specific events, episodes, or contexts. Semantic memory is about rules, categories, generalities, relations, and associations and can be context free. The forgetting of amnesics is characteristically episodic, whereas their semantic memory remains normal. The apparent time gradient in amnesics is due more to the nature of the interview, which focuses on episodic material for recent events and semantic material for remote events.

Kinsbourne and Wood (1975) reviewed the experimental work on amnesics that has been done using the Peterson-Brown paradigm. The basic results of their review suggest strongly that amnesics performed more poorly on the STM task than did normals, but the forgetting curve was parallel for amnesics, and for normals over an interval from 3 to 18 sec, there was no interaction between the groups. This indicates that the STM memory process is the same for both groups; they forget in the same fashion but differ simply in the amount that is forgotten. Also, the amount of interference interacts with retention time in the same way for both. Kinsbourne and Wood cite studies (H. Gardner, Boller, Moreines, & Butters, 1973; Wood & Kinsbourne, Note 1) in which subjects were told at retrieval time which category the words that they were required to learn were taken from, and retrieval was aided. Similar information at the time of learning had no effect. They concluded that "the balance of the evidence favors a retrieval explanation of the amnesic defect in short-term memory as measured in the Peterson and Peterson paradigm" (Kinsbourne & Wood, 1975, p. 277).

Conclusion

It seems clear that for those who wish still to hold to the distinction, STM and LTM share two important features: (a) Interference is a major determiner of forgetting in both, and (b) they both involve articulatory and semantic coding, although STM probably makes a greater use of the former. There remain differences: (a) Maintenance rehearsal can occur in STM but not in LTM, and (b) STM is ordinarily viewed as a port of entry into LTM for new memories or a temporary holding template until a permanent memory is formed. On the basis of the human data and observations, however, there is little support for the notion that STM is necessary for the firming of a physiological memory trace in the sense that Müller and Pilzecker (1900) originally posited or that the passage of time alone in STM allows the trace to consolidate, a conception that still prevails among animal physiological psychologists (Gold & McGaugh, 1975). Memories seem to be transferred to LTM by "mental" operations, such as the various forms of coding that can occur with the speed of the neural impulse, and the source of the forgetting seems to be at retrieval rather than storage.

This conclusion is shared by Isaacson and Pribram (1975), who, in writing a summary for their book on the hippocampus, say,

> any simple, long-term memory consolidation hypothesis of hippocampal function based on the initial findings with human subjects has become untenable in

the light of subsequent analysis. Unfortunately, this hypothesis is still held by the majority of people not actively involved in hippocampal research. (p. 43)

This is a conclusion similar to that of Weiskrantz and Warrington (1975) who say, "the amnesic deficit appears to be with mechanisms beyond the initial input into storage" (p. 413).

The quantity of surgical, chemical, and electrical intervention that has gone into animal research on this topic has not been possible with humans. It is time now to turn to animal research. There is, of course, no guarantee that human and animal memory processes are the same or even similar, but since Pavlov's studies on the dog, marked commonalities in learning have appeared, and it is not unreasonable to assume that the same will be true for memory.

Evidence From Animal Studies

General

Research on animal memory has been undertaken largely by physiological psychologists whose basic purpose has been to discover those physiological processes that accompany and probably are an essential part of learning. Since learning is a relatively permanent change in behavior, a basic assumption is that a physical structure must be involved. It is assumed that a time period is necessary for the structure to form, but it is also clear that there is memory almost instantaneously following stimulation. Therefore, some process intervening between stimulation and final structure has been assumed to be necessary. Hebb (1949) assumed that reverberatory electrical process could maintain the memory until the structure was formed. But most would agree that the electroconvulsive shock (ECS) studies convincingly demonstrate that this cannot be true. First of all, ECS sets up an electrical storm in the brain that totally overwhelms the kind of patterned electrical activity posited by Hebb. In addition, there is an almost total isoelectric period following ECS that does not permit patterned electrochemical neural transmission.

Hebb's is a two-process sequential model, with STM leading to and triggering the LTM. Two-process parallel theories have also been proposed in which both STM and LTM are initiated simultaneously at the time of learning, but the STM drops out after a critical period, leaving LTM to carry the memory. Gold and McGaugh (1975) have proposed a single-trace but two-process memory system. The trace is formed immediately at acquisition, but it fades and disappears unless some hormonal process follows that produces a memory fixation. Each of these alternatives is considered to be a form of consolidation theory in that each requires an initially fragile period for memory during which it is open to disruption or fading.

Evidence for STM–LTM

Duncan (1949) performed the classical study that showed that test performance was a function of the interval between a learning trial and the administration of ECS; the closer in time the ECS is to the learning event, the less the test performance. This function was called "the gradient of retrograde amnesia" and is one of the most well-established facts in the memory literature. In an influential article, Glickman (1961) reviewed the ECS literature and concluded that this gradient reflected the period for the consolidation of the learning trace. Although there was sporadic disagreement about this conclusion (see Coons & Miller, 1961; Lewis & Maher, 1965, 1966), it was vigorously reaffirmed by McGaugh (1966).

Methodologically, research moved from the use of multiple learning trials and multiple ECSs (Duncan, 1949) to single trials and single ECSs. Single learning trials were used because the age of the learning trace was made more determinate, and single ECSs were used because data showed that whereas several ECSs had marked punishing effects, one usually did not (Hudspeth, McGaugh, & Thomson, 1964). Experiment after experiment has been reported over the past decade in which a "time-dependent" amnesia has been found. The gradient varies with the intensity of the reinforcer used in acquisition (Mah & Albert, 1973), with the intensity and duration of the amnesic agent (Alpern & Mc-

Gaugh, 1968; Buckholtz & Bowman, 1972), and with other experimental conditions (Cherkin, 1969). Amnesia gradients have been produced with convulsant drugs (Pearlman, Sharpless, & Jarvik, 1961), hypothermia anesthetics (Alpern & Kimble, 1967), as well as electrical stimulation of the brain. The most widely accepted conclusion has been that the effect is on the storage of information and that at least indirectly, the gradient reflects the time for memories to pass from a fragile phase to a more permanent one (McGaugh & Dawson, 1971). Originally it was believed that ECS totally disrupted the consolidation process and destroyed the memory (Luttges & McGaugh, 1967), but now consolidation is believed to slow down and speed up depending on a number of experimental conditions (Mah & Albert, 1973), and the disruption may leave a partial memory that can later be reactivated and become fully consolidated and permanent (Cherkin, 1972; Kesner, 1973).

Additional support for the consolidation theory came from studies in which a block in protein synthesis in the brain coincided with a memory block (J. B. Flexner, Flexner, & Stellar, 1963; L. B. Flexner, Flexner, & Roberts, 1967). Various antibiotics injected into the brain or administered subcutaneously were shown to disrupt the formation of new protein by as much as 95%, and a memory impairment for recently acquired information was produced as well. Since, it was conjectured, a new structure mediating LTM must involve protein formation, a plausible mechanism for permanent memory seemed at hand. Further, injections of the protein synthesis inhibitor cycloheximide (CYCLO) before learning did not prevent normal initial acquisition nor did it interfere wth memory until several hours (3 to 6) had passed (Barondes & Cohen, 1968a, 1968b), by which time severe memory decrements had been observed. Thus, CYCLO seemed to be acting specifically on LTM while leaving STM intact. A similar conclusion may be drawn from data showing that memory is intact shortly after ECS (Geller & Jarvik, 1968) but disappears rapidly thereafter. Originally it was believed that ECS disrupted the consolidation of STM, but these later data are interpreted to mean that STM is left intact, and the interference is with LTM or with the transfer from STM to LTM.

The locations of physical sites in the brain for these effects have been pinpointed. Amnesia can be reduced if subseizure electrical current is administered to the hippocampus (Kesner & Connor, 1972, 1974), the caudate nucleus (Wilburn & Kesner, 1972), the substantia nigra (Routtenberg & Holtzman, 1973), the thalamus (Mahut, 1962), the midbrain reticular formation (Glickman, 1958), and the amygdala (McDonough & Kesner, 1971). (See McGaugh & Gold, 1976, for a review of these studies.) In each case the amnesia was time dependent. Bloch (1976) provides evidence for the notion that consolidation can be speeded up as well as slowed down. Time-dependent phenomena are again prevalent and time dependency is widely deemed to uniquely reflect the progressive formation of a new permanent memory structure and the fading or decay of an STM.

Evidence Against STM–LTM

Some data causing difficulties for the traditional consolidation of a STM into LTM are reviewed, and some of the newer theoretical notions are introduced. The negative data concerning the dichotomy between STM and LTM are treated more extensively than the positive, since that dichotomy is less widely accepted.

General

The supreme difficulty with the early consolidation theory (McGaugh, 1966) was that it failed to take sufficient notice of the learning–performance distinction (see Lewis, 1969). Thus, the behavioral deficits produced by amnesic agents administered at the time of learning were believed to be a direct reflection of the consolidation engram. Insufficient attention was paid to the many reasons why animals fail to perform other than that they no longer have any memory for what was learned. As a consequence, the RA gradient was believed to reflect the length of time during which STM remained open to disruption. This was generally believed to be a matter of

seconds or a few minutes at most (Glickman, 1961), and since the engram was a chemical–electrical–physical event encased in a relative constant physiological environment, its consolidation time was assumed to be relatively invariant. This latter assumption turned out to be far from the truth. Consolidation gradients had great variability from approximately 10 seconds (Chorover & Schiller, 1965) to 6 hours and more (Kopp, Bohdanecky, & Jarvik, 1966). It became necessary to consider the empirical RA gradient as reflecting only the period of susceptibility to disruption of the memory and not as the actual consolidation time. The reports that the consolidation time could be reduced to less than .5 sec (Lewis, Miller, & Misanin, 1968, 1969) if the animals were well familiarized with the learning environment before the introduction of footshock (FS) and ECS were often overlooked. (See Dawson & McGaugh, 1969b, for a failure to replicate; and Hinderliter, Smith, & Misanin, 1973; Jensen & Riccio, 1970; Miller, 1970; Riccio & Stikes, 1969, among others, for positive instances.) Lewis et al. (1969) suggested that learning occurred with the speed of the neural impulse and that ECS as ordinarily administered must be affecting a retrieval process, not a storage one. If amnesia is due to a retrieval failure, then RA gradients are not informative about consolidation time or memory fixation.

Attributing the action of an amnesic agent that is administered at the time of acquisition to a retrieval process has seemed illogical to some researchers. They prefer to believe that experimental agents acting at acquisition must be on acquisition processes and that only agents acting at the time of retrieval can be on retrieval processes. Many human cognitive psychologists, however, have long believed that interpolated learning has its effect on retrieval mechanism, even though the interpolated learning closely follows the original learning, and the test is much later. Interpolated learning may not work in the same fashion as an amnesic agent, but the point is that the proximity of the forgetting agent to original learning need not demand that any memory deficit be due to storage failure.

Reminder and Recovery

Another difficult problem for the storage failure point of view rose when the "reminder" studies began to appear (Koppenaal, Jagoda, & Cruce, 1967; Lewis et al. 1968). The basic assumption of these studies was that ECS served to inhibit or block the retrieval of memory rather than disrupt its formation. This approach assumed that memories endured through ECS, but access to them was temporarily lost. If so, then a reminder—a portion of the original learning situation—should return them to expression. Miller and Springer (1972) explored the reminder effect in detail and found it to be independent of the ECS-reminder interval or the reminder-test interval. Also, it occurs in appetitive situations as well as those of avoidance.

A significant extension of the reminder effect was reported by Quartermain, McEwen, and Azmitia (1970). They used CYCLO to bring about a deep inhibition of protein synthesis and an amnesia. Nevertheless, a reminder stimulus produced a memory return, which suggested strongly that whatever the effect the inhibitors of protein synthesis had on memory, it did not block its formation. It must be understood that the effect of any drug on memory depends on many conditions other than the drug itself, such as the strength of the memory and the test conditions. Nevertheless, the role protein synthesis inhibition plays in memory formation is probably not great, if it plays a role at all. At least memories can be fully recovered even when protein synthesis is inhibited over 90% (Quartermain, 1976).

Lost memories can also be returned to expression pharmacologically. Perhaps the first clear demonstration of this was the Braun, Meyer, and Meyer (1966) study in which a visual discrimination was lost following extirpation of the posterior cortex of the brain but was relearned rapidly, compared to controls, on injections of DL-amphetamine. Adams, Hoblit, and Sutker (1969) have returned memory to expression following ECS with injections of physostigmine. Roberts, Flexner, and Flexner (1970), Serota, Roberts, and Flexner (1972), and

Botwinick and Quartermain (1974) also have induced memory recovery pharmacologically following a memory lapse due to protein synthesis inhibition. Perhaps the most important recoveries have been reported by Rigter and Van Riezen (1975). (See Barraco & Stettner, 1976, and Meyer & Beattie, 1977, for more detailed reviews.) The recovery of memory following protein synthesis inhibition shows that there could not have been a permanent loss of LTM even though memory seems to disappear approximately 6 hours after acquisition.

Interestingly, protein synthesis inhibition compounds also reduce the amount of available norepinephrine (NE) and other biogenic amines, and Serota et al. (1972) proposed that a NE deficiency lies behind the amnesias. If so, Quartermain (1976) has shown that the effect is on memory retrieval and not on its formation in that he has found a spontaneous recovery of a T-maze memory for which amnesia had been produced by the inhibition of NE synthesis. Barraco and Stettner (1976) conclude, "we believe that the leading hypothesis at present should be that antibiotics block or impair retrieval processes providing access to memories for specific aspects of training" (p. 271). They add that puromycin may be an exception to this conclusion.

The fact that memory either recovers spontaneously, recovers on reminder stimulation, or is recovered pharmacologically indicates that the amnesias are not due to the failure of memory formation, either short term or long term. Reminder stimulation has not been tried in all situations, and therefore this conclusion must remain tentative, but the body of evidence on amnesia strongly suggests that a retrieval mechanism is involved and that a multiple-stage memory formation theory is not necessary for currently existing facts.

Cue-Dependent Amnesia

Another block of evidence that runs counter to a two-process memory conception comes from studies on cue-dependent amnesia. These studies suggest a different variety of memory processing of a cognitive type (Lewis, 1976), which is discussed later. The first cue-dependent amnesia experiment was that of Misanin, Miller, and Lewis (1968; but see Schneider & Sherman, 1968, for a similar experiment with a different point of view). First, the animals learned an avoidance response cued to a tone. Twenty-four hours following learning, after any consolidation process should have been completed, the animals were returned to the learning apparatus, the tone was presented, and it was immediately followed by ECS. Twenty-four hours after treatment, the animals were returned for testing, and those that had received ECS in the presence of the learning cue were amnesic. The implications of this study were considerable. It suggested that the age of the memory was not an essential determiner of whether or not amnesia would be produced by ECS and that amnesia was not dependent on a memory's early fragility. An immediate implication was that the RA gradient was not conclusive evidence of consolidation failure or even of a susceptibility to storage disruption (Lewis, 1969). The Misanin et al. (1968) study also showed that ECS had to be given in the presence of a learning cue from the learning situation for memory to be disrupted, and for this reason the effect has been called cue-dependent amnesia.

Misanin et al. (1968) speculated that the presence of the cue served to reinstate the memory and that the simultaneous presence of ECS with the memory served to inhibit or block the later expression of that memory. In other words, the memory must be evoked for it to be blocked by ECS and presumably by other amnesic agents. This reasoning led them to make a distinction between active and inactive memories. Learning always occurs in the presence of specific cues and a contextual environment, and to the extent that these cues and context occur again the memory will be reinstated. Thus, memories are active under at least two conditions: (a) during original learning and (b) during reinstatement. Under both conditions the memories are open to disruption by ECS. Since memories that are not reinstated are least disturbed by ECS, they were believed to be

in a different state, a state of inactivity or passivity. Active memories could be disrupted; inactive memories could not.

This experiment has been essentially replicated by many researchers (although see Dawson & McGaugh, 1969a, for a failure to replicate). Davis and his colleagues were simultaneously reporting a similar finding. Davis and Klinger (1969) extended the interval over which potassium chloride, puromycin, and acetoxycycloheximide could produce amnesia by leaving their subjects in the conditioning chamber. Confinement in the training apparatus extended the effective amnesia interval in experiments by Davis and Hirtzel (1970) and Potts (1971). Schneider and Sherman (1968) reported a similar finding. Meyer and his students (Howard, Glendenning, & Meyer, 1974; Howard & Meyer, 1971; Robbins & Meyer, 1970), using a multiple-choice maze, gave animals three different successive problems, each associated with its characteristic cues. They reported amnesia for problems whose cues were paired with ECS, regardless of the age of the memories, which were frequently 3 weeks old. A similar cue-dependent amnesia has been demonstrated many times (DeVietti & Holliday, 1972; Gordon & Spear, 1973a, 1973b; Lewis & Bregman, 1973; Lewis, Bregman, & Mahan, 1972). Finally, DeVietti and Kirkpatrick (1976) and Gordon (1977b) have shown a typically RA gradient in a cue-dependent amnesia situation, which confirms the inference that time-dependent vulnerability of memory is not uniquely associated with storage. Of course, the time dependency of reactivated memories may be due to a different mechanism than for a new memory, but it is unparsimonious to believe this until data supporting two mechanisms are presented. Gordon (1977a) has shown that the reactivated gradient is shorter than a new gradient, but he does not believe this finding is support for two different mechanisms.

The reactivation situation has also been used to show the kind of memory enhancement that has previously been demonstrated when certain drugs are administered soon after a learning experience. Gordon and Spear (1973b) showed that strychnine sulfate could enhance a well-aged memory if and only if the memory were reactivated immediately prior to drug administration, and Gordon (1977b) has shown that this effect is time dependent. Finally, Landfield, McGaugh, & Tusa (1972) found an electroencephalogram (EEG) Θ wave that appeared immediately following learning, and they interpreted the Θ as an indicator of a memory consolidation process. Nicholas, Galbraith, and Lewis (1976) confirmed their empirical finding but found the Θ to occur also upon reactivation of the memory. They interpreted the Θ as an indicator of memory activity.

Conclusion

All of the data presented in this section argues strongly against the standard two-process sequential model in which STM precedes LTM and is a necessary holding template for the LTM. But there are other forms of a two-process model (see McGaugh & Dawson, 1971). For example, STM and LTM could both be initiated at the onset of learning and could parallel each other until STM drops out, while LTM continues. This is a parallel rather than a sequential two-process theory. The basic point being made here, however, is that memories can be recovered following amnesia, regardless of whether the amnesia is conceived of as a failure of STM, LTM, or the transfer between the two. There are, of course, other conceptions of memory formation than have been considered here, but they are variants on a general theme. The theme is that following an experience, new memories are fragile and will decay unless they are either held by rehearsal in STM or transformed into something more permanent. This transference into LTM takes place typically as a function of time or of time and something else, for example, a nonspecific physiological response, usually adrenergic (Barondes & Cohen, 1968a; Gold & McGaugh, 1975). Various arguments and data have been cited to show the inadequacy of these conceptions of an initially fragile STM or LTM. Both intuitive experience and experimental data

indicate that memories are formed almost instantaneously following an experience and that these memories are relatively permanent, are surprisingly resistant to simply decay, and seem to yield only to forms of interference and competition that affect retrieval.

Several powerful guns have recently fired at the consolidation–storage hypothesis. Weiskrantz and Warrington (1975) say:

> But the lack of evidence for a defect of "consolidation" or long-term "storage" in animals is no longer an embarrassment, because that hypothesis appears to be no longer able to account for the amnesic defect in man. (p. 425)

Meyer and Beattie (1977) state:

> However, the argument that interventions which have time-dependent actions upon newly-formed habits produce their effects because they interfere with labile traces has been tested, and found wanting. The argument that interventions which produce complete and lasting impairments of retention must therefore have interfered with memory formation is not now, nor has it been for almost a decade, worthy of serious belief. (p. 154)

Apparently, memory can be formed within a time span of less than a second, and it is reasonable to believe that this may be typical of most memories. There is presently no reason to believe that a new memory is any more fragile than an established one. A physiological process must be found that accommodates such a brief memory formation time span, and we have already seen that reverberating neural firing is not the answer. The synthesis of new protein also probably does not provide a mechanism for the relative immediate formation of memory. Squire (1975) proposes that under the simplest of circumstances, at least 1 minute is required for the synthesis of protein and its transport to a synaptic site. A longer interval would be required if the prior synthesis of mRNA were necessary. This does not deny a role for protein synthesis and the growth of synaptic knobs (Lynch, Deadwyler, & Cotman, 1973; Rutledge, 1976) as a support for frequently evoked memories, but it does deny that these mechanisms are necessary for the formation of permanent memories.

The possibility that learning is due to changes in synaptic neurotransmission remains strong, but a long time-dependent structural change is not necessarily required for such changes to take place. They can, in fact, be rapid, almost as rapid as the elaboration of chemicals at the synapse or the combination of protein fragments or the addition of a carbohydrate to a neural membrane protein, forming a glycoprotein (Bogoch, 1968, 1973). Kety (1970) has proposed a model that suggests how biogenic amines could participate in the learning process. He proposes that novel or surprising stimuli cause the release of the appropriate biogenic amines throughout the central nervous system and that these amines serve to increase firing probability in neural systems that are active and perhaps decrease activity in others. A specific pattern of neural firing then comes under the command of the environmental events existing at the time. The pattern of neural firing is the memory, and it is reactivated when the stimuli are presented again.

Active—Inactive Memory Distinction

Active memory (AM) is considered as a changing subset of all permanent memories possessed by an organism. At any given time many of the permanent memories, which have the potential for being active, are in a relatively inactive state and have little effect on current behavior. A rat, for example, who has learned a complicated T maze has the potential to perform correctly in that maze even though, at a given moment, it is eating from a trough in a living cage in a room that is different from the one in which the learning occurred. Also, a memory may be active without having an observed effect on current behavior. A rat may be in the start box of the T maze and remain stationary for any one of a number of reasons (e.g., low motivation, sickness) other than the failure to reactivate a memory. Although we speak of active memory (AM) and inactive memory (IM), different stores or locations are not implied; it is doubtful that AM occupies a specific site or sites in the brain. It seems

better at present to conceive of AM as a patterned state of neural firing that is not specifically localized, although different memories that are active undoubtedly reflect different densities of firings in different parts of the brain. It is likely that each evocation of an AM will be slightly different from each other evocation, although there must be considerable similarities, or the memories could never be identified as the same (John, 1972).

In each human, AM is intuitively apparent; "immediate consciousness" is a part of it. In lower animals AM is not intuitively apparent except through anthropomorphic analogy. Positing an entity such as AM for rats is scientifically dangerous and certainly far outside the standard stimulus–response (S–R) tradition. Nevertheless, there are precedents for the application of cognitive concepts to rats and for their experimental manipulation. Tolman (e.g., 1932, 1949) was a biologically oriented psychologist whose use of behavioral-based but cognitive concepts such as "purpose," "cognitive maps," and "memorial lore" makes interesting reading today. Although Tolman did not use the concept of AM, his "vicarious-trial-and-error" (VTE) is instructively similar. For Tolman, VTE was most apparent at a choice point and was evidenced by the back-and-forth movement of a rat's head before it committed itself to one of the alternatives. In Tolman's view, the rat was weighing the alternatives, cognitively trying to arrive at the correct decision. He was able to quantify the VTE to some extent. The amount of VTE was a direct function of the difficulty of the problem. For easy discriminations VTEing occurred during the early trials and dropped out as learning proceded. For more difficult problems there was little early VTEing; the greatest amount of VTEing occurred just before and at the time of the solution to the problem and continued some time after. For the most difficult discrimination, VTEing was never greatly reduced after solution, even with a great amount of overlearning. Apparently, the animal had to continue to weigh alternatives on each trial. It may well be that VTE can be used as one index of AM.

Human Conceptions

The concept of STM as an entry point for new memories and their transfer to LTM goes back at least to James (1890). However, the concept of AM, which is simply an aspect of the total memory process into which memory can both flow from IM and serve as a point of entry for new memories, is, seemingly, new. The first mention of an active–inactive distinction may well have been made by Posner (1967), who proposed further that interference and forgetting could take place in AM. He briefly summarizes an experiment that was negative to the concept of AM as an arena of forgetting, but he developed the conception of AM in more detail later (Posner, 1973). Certainly, as is discussed later, AM as a process in which forgetting occurs needs much more experimental testing.

The distinction between AM and IM also developed through the theorizing of Atkinson and Shriffin. In their earlier treatment (1968) they still held to the standard notion of STM that served only as a point of entry and processing of new learning. In Atkinson and Shriffin (1971), however, they conceived of STM as a processing point for both new learning and for retrieved memory from LTM. For them, AM was the conscious activity of memories, regardless of whether they were old or new. It was the control center for rehearsal, coding, and imagining, for all cognitive activity. This is an important theoretical treatment of memory with a sophisticated discussion of AM.

A similar treatment of AM as the site for the processing of retrieved memories as well as for new learning is that of Craik and Lockhart (1972). They suggested that different levels of processing occurred in AM with the deeper semantic levels laying down more effective and retrievable traces. Baddeley and Hitch (1974) also have conceived of AM (working memory) for both storage and further processing of retrieved memories. They show that several cognitive processes (e.g., free recall) are made more difficult with additional simultaneous memory activity. This is because of the limited capacity of AM, which as the "focus of attention"

can perform only a circumscribed set of processes at one time. Bjork (1975) has presented a similar conception of AM, as have Shiffrin (1976) and Bower (1975), among others. Thus, a concept of AM as a central processor of both new learning and of retrieved memories now exists. It is similar to conscious awareness, attention, and even thinking as far as humans are concerned.

Animal Conceptions

For lower animals, the first distinction between AM and IM apparently was made by Misanin et al. (1968), who showed that rats could be made amnesic for old memories if, they hypothesized, the memories were active at the time of ECS, whereas IMs were relatively immune to the amnesic treatment. This distinction has been explored further by Lewis (1969, 1976), Spear (1976), and by Lett (1978). These theoretical and experimental developments are considered in the next sections.

Entry Into Active Memory

Two kinds of events can be distinguished for AM by their route of entry into AM. First, memories can be created anew from external stimulation and, second, memories may be reactivated later when these stimuli are again present. Because entry into AM as a new memory or reentry as an old one usually entails some processing of the sort discussed later, it is impossible to separate entirely the new entry function from the other, and some overlapping of these functions cannot be avoided.

New Memories

Each stimulus activating a receptor is converted into neural impulses and under normal conditions, occasions brain activity that consists of a representation of the stimulus complex, and that representation is one example of an AM. These representations simultaneously contain information about sensory mode or modes, context, temporal arrangements, color, shape, and other properties and relationships among stimuli. In total, these make up the attributes of memory (Underwood, 1969).

Memory is considered to be active for animals whenever the organism is alert and patterned stimuli impinge on the sensorium. Under these conditions it is assumed that the organism is either learning or retrieving. (See Thistlethwaite, 1951, for a review of the latent learning literature, and see Spear, 1973, for a discussion of the role of the stimulus and the context upon retrieval.) Reinforcement is not necessary for learning, but a reinforcer does serve to focus attention by pointing out to the organism that something specific is to be learned. If there is no reinforcement, they will be learning anyway, but it will be more difficult for the experimenter to discover what has been learned. This also means that a nonlearning control group for those who are searching for the physiological correlations of learning cannot be simply one for which reinforcement is not administered. After the organism has learned, the stimuli that were salient (significant, noticed, reinforced, emphasized) during learning can be presented again, and the memory of the learning will occur. Since the experimenter's notion of what the salient cues are may not also be those of the animal, representation (or reinstatement) of the cues may not always reactivate. A distinction between nominal and functional cues is important.

It would be helpful to determine the simplest possible memory. Learning is frequently considered to be an association of two or more items, stimuli, or events, but it becomes difficult at times even to know when two items are present. If, for example, the stimulus object is a square, the top half of which is one color and the bottom is another color, is the stimulus integral or separable (see, e.g., Blough, 1972; Gardner, 1970; Leith & Mahi, 1977)? Nevertheless, items vary in complexity, and it is possible to conceive of the simplest item along a complexity dimension. This kind of primitive conception will have to do for analytical purposes at present. An item of learning will be the conceptual simple unit. When items are related, an association is said to be formed.

It is assumed here that new simple memories have two important properties, properties that will appear startling to some: (a) They are formed almost instantaneously, and (b) they are relatively permanent. Neither one of these properties can be proved empirically, at least at present. Still, they are not empty properties, for they have consequences that can, at least to some degree, be empirically tested (see Lewis, 1969, 1976; Miller & Springer, 1973).

There are two fundamental sets of data that seem to contradict the notion that memory formation can be almost instantaneous. First is the retrograde amnesia gradient which implies that a newly formed memory becomes less susceptible to disruption over time, which, as we have already shown, requires reinterpretation, and the second is the typical learning curve that implies a gradual improvement over trials.

The learning curve seems to require an incremental process, proceeding bit by bit to an asymptote. But this empirical curve must be separated from the underlying theoretical process causing the curve. Although Hull (1943) assumed underlying incremental strength in habit for the learning curve, Guthrie (1935) did not. Guthrie assumed that learning occurred all at once and immediately whenever a response occurred in the presence of a stimulus. Learning was thus a one-trial affair. The incremental learning curve arose because the stimuli varied from trial to trial, and typically a number of trials were required for the response to become attached to a sufficient number of stimuli. Estes (1950) has formalized this assumption, which remains a common and important one (Bower, 1975). Similar one-experience learning assumptions have been made by those studying amnesia (Irwin, Banuazizi, Kalsner, & Curtis, 1969; Lewis, 1976). Recent treatments of learning (see Bower, 1975) are not concerned with the relationship between a stimulus and a response but with the relationship between a stimulus and the contents of AM. Since the contents of AM are at least as variable as the stimuli, another reason for an incremental learning curve is present, even assuming immediate and permanent learning. The point here is that neither the incremental learning curve nor the decremental amnesia gradient requires a theoretical growth or decay process.

Much learning, of course, involves considerable complexity and many items and therefore requires more experience and more trials than does the simplest learning. Obviously, it will take longer to learn a T maze with 10 choice points than a maze with only 1. Multiple-unit learning requires that each unit and the relationship among the units be learned. Complexities of this type have led biological researchers to employ the simplest types of learning situations possible. Typically, two have been selected: Pavlovian conditioning and single-trial passive avoidance learning. Each of these probably still remains complex, relative to what the simplest possible learning situation could be. Even so, the evidence suggests that an animal can learn quickly that a footshock (FS) has occurred in the passive avoidance situation (Lewis et al., 1968) as shown by its avoidance of the FS location on the next trial. The animal may learn only that an FS has occurred, or it may learn also that the source of the FS was the grids on the floor, something about the intensity and quality of the shock, what happened before the shock, and what happened after the shock, all of which will greatly aid in the retrieval of the memory about FS. Such further learning will take more time than the simple learning that an FS occurred, but each item of learning is assumed to occur rapidly and to be permanent.

Established Memories

Perhaps equal in importance to learning, as a source for AM, although frequently neglected, is IM. It is clear that memories have the potential to return to expression long after original learning. It is argued that all contents in AM share similar properties, that is, a reactivated old memory is similar in many respects to a newly created AM.

When stimuli from original learning impinge on sense organs, and the organism attends to them, perhaps as indexed by the

orienting response, a characteristic pattern of neural firing is reinstated, and in some fashion this reinstated firing represents the memory. This kind of conceptualization is different from a traditional S–R paradigm. Adams and Lewis (1962), for example, in the traditional format, proposed that ECS served as an unconditioned stimulus (US) in the classical conditioning sense and that the conditioned competing response evoked by ECS replaced those that had previously been learned in an avoidance situation. The replacement of the original learning appeared to be an amnesia. This interpretation was soon found to be inadequate for the single-ECS situations to which researchers turned because no evidence of a conditioned convulsion could be detected with only one ECS administration (Paolino, Quartermain, & Miller, 1966). Once freed from the bondage of thinking in terms of traditional peripheral responses, it was possible to think in terms of events and memorial representations. Whereas Adams and Lewis thought of competing responses, and Lewis and Maher (1965) proposed the inhibition of peripheral responses, Misanin et al. (1968), in their cue-dependent amnesia experiment, thought of the inhibition of memories.

Because of the imprecise relationship—to the experimenter—of memories to stimuli and responses, it is not always possible for the experimenter to present the exact cues that will reinstate the memory. Garcia and his colleagues (e.g., Garcia & Koelling, 1966; Garcia, Kovner, & Green, 1970; Garcia, McGowan, Erwin, & Koeuer, 1968) have forcefully illustrated that all stimuli are not equivalent in their ability to initiate memories. Ordinarily, for an experimenter to test successfully for memory, either the associated events had to occur in close temporal contiguity or the animal had to maintain a physical orientation to the stimuli (Grice, 1948; Hunter, 1913; Spence, 1947). But Garcia and his colleagues showed that if the stimulus is a novel taste and the response is illness, then even hours may separate the two and still the rat will avoid the novel stimulus the next time it is presented. This phenomena is difficult to formulate in standard learning terms because of the long time interval between the novel taste and the induced illness. Nor does the concept of preparedness (e.g., Seligman, 1970) provide a satisfactory answer to the basic question: How does the rat bridge the gap between the novel stimulus and the sickness? There are several possibilities:

1. Novel tastes create a pattern of neural firing that endures long enough so that the firing pattern from the illness becomes contiguous with that from the taste, and an association is formed. If so, the taste pattern has to subsist through hours of other experiences and thus many other patterns of firing.
2. Illness evokes past taste sensations and novel tastes are the most salient with illness, and thus there is a contiguity between the memory of taste and the illness.
3. Taste aversions in the rat are a special kind of learning for which the animal is innately prepared, and it is simply outside the ordinary laws of learning.
4. Illness creates a phobic reaction to novel tastes in general (Mitchell, Scott, & Mitchell, 1977).

Data that distinguish among these alternatives are not yet available, but if learning is involved, it is difficult to avoid the necessity of a relative contiguity between the associated events, which means that illness probably reactivates the memory of the novel taste, and the association is formed in AM between the reactivated event and the new event (Lett, 1973, 1978).

In summary, it seems that memories not only enter AM from IM but are produced from the external world in a new state. If so, then a very important question is to what extent are reactivated memories like new memories?

Comparison Between New and Established Memories

It has frequently been shown that the disruption of new memories by amnesic agents is time dependent, producing the RA gradient. DeVietti and Kirkpatrick (1976) showed a similar time dependency for reactivated memories, as has Gordon (1977a, 1977b), although Gordon found the effective gradient

shorter for the reactivated memory than for the new one. Because gradients have been found to be so tremendously variable (see Chorover, 1976) even for new memories, this difference is probably not significant. Time-dependent gradients have been produced for reactivated memories by electrical stimulation of the brain through implanted electrodes (DeVietti & Kirkpatrick, 1976), by strychnine sulphate (Gordon, 1977b), and by other competing memories (Gordon & Spear, 1973b). An enhancement gradient for a reactivated memory has also been reported by DeVietti, Conger, and Kirkpatrick (1977).

A retrieved memory may be different from what it was during original learning because some part of it may be obscured due to competition and interference, or it may have become a part of a different memory complex from when it entered memory. Because memories are dynamic, it is unlikely that a retrieved memory will ever be identical to the original, and the very act of retrieval will produce a change. It is also likely that there will always remain enough of this original memory in what is retrieved for the two to be recognizably similar.

As suggested, the process of retrieval itself has an effect on the memory, but it is far from clear what this effect is. Cherkin (1970) has argued that retrieval amounts to additional learning and that the effect of a reminder cue is the same as the effect of an original learning trial. The effects of reactivation on the memory have still to be worked out in detail, but sufficient data exist which show that a learning trial and reinstatement or reactivation trial are not the same (Lewis & Nicholas, 1973; Gordon & Spear, 1973a). Reinstatement is the presentation of the cues that were present during learning but without a reinforcer. Clearly, this is the experimental operation for extinction as well as reinstatement, and extinction is not an additional learning trial. There are other differences between the two, however. For extinction, many trials are repeated with a fairly short intertrial interval, and response decrement increases over trials. For reinstatement, typically, only one trial is given, or if two or three trials are given (Spear & Parsons, 1976), they are widely distributed. It is probable that similar memorial processes are involved for both procedures, but there is a difference in their consequences. In the case of extinction, the successive failures to reinforce indicate that the previously reinforced cues and the associated behavior are no longer significant. Orienting and other alerting behavior are reduced. For reinstatement, presentation of cues without reinforcement is unexpected, producing orienting and rehearsal. The status of the cues as indicators of reinforcement is indeterminate. Also, because reinstatement usually occurs at a long interval following the last learning trial, there is reason to suppose that the memory has been degraded to some extent at the time of the representation of the cues. The degradation can be a function of the stimulus change and memory interference that inevitably occur with time (Campbell & Jaynes, 1966), or it can be due to experimental procedures such as the administration of an amnesic agent.

A first guess is that reactivation has a much stronger effect, at least in multitrial situations, than does a single additional learning trial at the time of learning, that is, a memory return from reactivation following forgetting or amnesia is much more dramatic than the administration of an additional trial during learning (Campbell & Jaynes, 1966; Lewis & Nicholas, 1973; Spear & Parsons, 1976). This suggests that the strong reactivation effect is somehow a function of the partial memory degradation. There is probably a definitional problem in distinguishing reactivation from the effects of distributed learning, although both could be on the same dimension as the beneficial effects of distributed practice (Hill & Spear, 1962; Hintzman, 1974) due to the recovery from the memory degradation that occurs during the intertrial interval.

Processing in Active Memory

In this section some of the processing operations performed on both old and new memories in AM are discussed. Since some processing necessarily occurs both when new

memories are formed and when established memories reenter AM, process similarities are inevitable.

There are four functions of AM that are tentatively defined as follows: (a) to register new inputs, to note that an event has occurred; (b) to associate two or more new inputs to each other, as illustrated by classical conditioning, the first phase of sensory preconditioning (Brogden, 1939), and when "one thing leads to another" (Tolman, 1949) as in latent learning; (c) to associate new learning with already established and reactivated memories. It is assumed that when the established memory is reactivated in AM simultaneously with the learning, an association is formed. The second phase of sensory preconditioning is an illustration of this process. The first conditioned stimulus (CS_1) is paired with the second (CS_2) a number of times, and an association is established (first phase). The CS_2 is then paired with a US (second phase), bringing about an association of the US with the established memory of CS_1 and CS_2. Of course, the degree of association and coding will vary with the different components. A more common example of an association between a new and an established memory is that of secondary reinforcement. CS_1 is first paired with a US. Then the established properties of CS_1 are paired with a new stimulus CS_2, producing a CS_1–CS_2 association; and (d) to associate established memories with other established memories. This occurs when two old memories are reactivated simultaneously (Hearst & Peterson, 1973; Solomon & Turner, 1962). Such studies are commonly performed to show the effect of classical conditioning on instrumental responding and to illustrate how two established memories can affect each other even though they do not share a common response.

AM Functions 1 and 2 have been the object of much laboratory research, but Functions 3 and 4 have not received the attention they deserve, and they are probably more interesting. They involve the study of memory structures and systems. Certainly these functions will soon attract the attention of inventive researchers.

Postresponse Events

In traditional S–R theory a learning trial was considered completed when the designated response was over and reinforcement was administered. It is now known that there are post response events that are extremely important to the later use of the memory. Duncan (1949) made this point in his classical study of amnesia when he demonstrated that a gradient was a function of the time interval from the termination of the learned response to the administration of ECS. Lewis and his colleagues (see Lewis et al., 1968) have long argued that the interval immediately following a learning trial is used by the organism for cognitive processes that aid in the retrieval of the memory. Tolman (1949) gave an illustration of a postresponse cognitive process when he described an experiment by Hudson who found no avoidance learning to electric shock if the apparatus cues surrounding the shock were removed immediately following treatment. If, however, the cue remained so that the animal could look back "to see what happened," as Tolman put it, it readily learned to avoid. Essential aspects of this study have been replicated by Keith-Lucas and Guttman (1975).

Wagner, Rudy, and Whitlow (1973) showed that if an unexpected stimulus was presented following a learning trial, retention was impeded, and the interference was a function of the time interval between the learning experience and the unexpected stimulus. They believed that adequate learning required a posttrial rehearsal process that was prevented by the unexpected stimulus. Similar data have been presented by Terry and Wagner (1975) and Miller, Misanin, and Lewis (1969). Also for humans, Waugh and Norman (1965) showed that any new item can displace an earlier one if it is unexpected but that there was no interference from highly predictable and redundant items. Bartus and Johnson (see Bartus & LeVere, 1976), with monkey subjects, found that the interference during postresponse processing was a function of the similarity of the postresponse stimulus to the relevant learning stimulus, indicating a retroactive interference

phenomenon. These studies emphasize the importance of postresponse processing for adequate retrieval to occur. If this process is interrupted by surprising stimuli, by stimuli that make it difficult to discriminate the important learning stimuli, or by the electrical stimulation of the brain, retrieval is interfered with.

What is the nature of the postresponse rehearsal process? It would be convenient to conceive of it as a sort of perseveration of the original input or a sort of continuing repetition. Although something of this sort may occur, it is also likely that a more active and cognitive activity takes place. Lewis (1976) hypothesized two phases of learning. First, the organism registers that the event has happened—an FS, the intrusion of another organism, and so forth. Second, there is a form of interpretation of the event, a relating of the new to the established, that is, that it was painful and came from the floor or that it was female and in heat, and so forth. It is this elaboration, following the initial registration of the occurrence of the event, that occupies the postresponse time. (See Bower, 1972, 1975; Everett & Corson, 1973; and Greeno, 1970, for a possibly similar notion.) This is certainly what Tulving (1970) had in mind when he said, referring to humans:

> When a to-be-remembered unit is stored, some ancillary information about it is also stored with it. The storage of this ancillary information represents what is referred to as "coding." When some of this ancillary information (or the "code" of the to-be-remembered unit) is available at the time of attempted recall, the code serves as a retrieval cue. (p. 8)

Sara, David-Remacle, and Lefevre (1975) also argue for a "core" memory that is blocked by ECS and that can be elaborated by further experience, returning the core memory to expression. Azmitia, McEwen, and Quartermain (1972) found no amnesia if the subjects recovered from ECS in the training environment. They also argue for a core memory that is not disrupted by ECS and for a further learning stage, "which may involve integration of the experimental cues associated with the core trace with the animal's existing memory repertoire of accessible experiences" (p. 855). It may also be similar to the distinction between item and order information (Murdock, 1974). The animal is trying to "make sense" out of what happened, to relate the event to the remainder of its experience. The distinction is perhaps also similar to that between the data and the address for the data in computer language. If no address is given for the data, their retrieval will be difficult indeed.

If the target event is commonplace and has been repeated many times, then the memory elaboration need not occur, other than to record that the event has happened again. The event is immediately recognized as familiar and readily fits the memories (event representations) already in existence. If the target event is surprising or unexpected, the rehearsal process is engaged to interpret the event, to fit it into an existing memory system, or to recognize and differentiate the new event from other events. These elaborative processes (rehearsal) take place in AM, and since one of the important characteristics of AM is its severely limited capacity, processing of other temporally near events must be curtailed, and therefore their retrieval will be reduced (Wagner et al., 1973). These postresponse processes involve new and rapidly formed memories that serve the purpose of aiding the retrieval of the memory of the target event. The more the event is distinctive and becomes part of an already existing memory system, the better is its retrieval (see Lewis, 1976). If postresponse stimuli are presented that are similar to the learning stimuli, retrieval is made difficult through interference (Bartus & Levere, 1976).

Preresponse Events

It is proposed that conditions existing prior to the target event also have an effect on memory processing. Lewis et al. (1968) showed that animals familiarized with the apparatus in which they were to receive FS and immediate ECS demonstrated reduced amnesia. They reasoned that this was because the properties and location of the FS were easy to distinguish in the familiarized en-

vironment, and the meaningful coding of the FS was simple and took less than .5 sec (Lewis, 1976). If an animal already knows "what leads to what" (Tolman, 1932) in an environment, the introduction of a new salient or biologically relevant stimulus is quickly assimilated (automatic search) into the "cognitive map" of the environment, leading to appropriate behavior. Lubow, Rifkin, and Alek (1976) present interesting data on this effect. They show easy learning when a novel stimulus is used in a familiar environment and when a familiar stimulus is used in a novel environment as compared to using a novel stimulus in a novel environment or a familiar stimulus in a familiar environment. This suggests that a familiar stimulus will not result in amnesia in a new environment.

Whether or not animals learn without reinforcement is no longer an issue; they do. Latent learning is discussed here as a form of familiarization with the environment that facilitates further learning when the experimental stimulus is introduced—either the reinforcer (Blodgett, 1929) or the FS (Lewis et al., 1968). The preresponse familiarization builds a cognitive structure into which the new event is readily assimilated. The importance of these preresponse events is further illustrated by Mitchell et al. (1977) who found that a long-delayed poison-induced aversion to novel visual stimuli in rats can be readily obtained if the animals have had extensive familiarization (habituation) with the environment prior to the introduction of the novel stimulus. Without the familiarization, visual stimuli do not readily become capable of eliciting delayed aversions.

The notion that prior familiarization facilitates new learning in the familiarized situation has human counterparts. Shiffrin (1976), who also takes the point of view that forgetting is a retrieval problem, discusses the limitations on retrieval posed by the limited capacities of AM. He says that search in AM can be either automatic or controlled.

> Automatic search occurs when a stimulus or set of stimuli is sufficiently trained with respect to a given background (or distractor stimuli) that it becomes associated with an automatic attention response that will cause the given stimulus to be searched or considered first, before competing stimuli. (p. 220)

Again, he says:

> Automatic search is facilitated by targets having distinctive simple general physical characteristics relative to the background (distinctive color, shape, orientation, location, etc.). In addition, automatic search can develop if any set of targets is trained sufficiently long with respect to a given background. (p. 224)

A similar notion has been presented by Sperber, Greenfield, and House (1973), who argue that an adequate representation or effective code of an item is more difficult to form when the item is unfamiliar than when it is familiar and already coded.

In summary, the characteristics of AM, particularly its limitations on the number of items that can be active at any time, have a great effect on what is retrieved. New learning will be easily retrieved if it occurs in an environment that is familiar and that provides, through existing memory representations of the environment, a memory context into which the new event may fit. If a new stimulus is introduced into a less familiar environment, more time will be required to fit the new learning into some existing memory context. This may be better conceived of as a coding process rather than rehearsal, which implies a mere repetition or holding of memory items. Tests of this distinction can be made by manipulating both prelearning and postlearning environments.

Loss of Accessibility to AM

Perceptual Stimuli

Since stimuli are almost always impinging on receptors, there is an immense amount of incoming information. Each head movement brings new stimulation and new information. It is clear that a great part of this new information, if it is different in any respect from memories, is almost immediately lost, and a basic question concerns the mechanism of the loss. There are two primary alternatives:

1. The consolidation process, which maintains that new information either fades or is open to destruction for a period of time following its formation and that incoming information destroys current newly formed old information (McGaugh, 1966; McGaugh & Gold, 1976; Müller & Pilzecker, 1900). At the receptor image level, there is a basis for fading images (Sternberg, 1966), but this notion has been rejected previously in this article as the explanation for the forgetting of memories.

2. The interference hypothesis, which maintains that the new items interfere with each other based on various properties such as similarity and the difficulty in discriminating among many similar items as the organism moves through the environment. This seems a reasonable explanation even for the forgetting of the multitude of perceptual items and is identical to the retrieval hypothesis.

Active Memory Forgetting

Until 1969, all studies of experimental amnesia had at least three properties in common. One was the presentation of cues for new learning, the second was the administration of a reinforcer, usually negative, and the third was the administration of the amnesic agent. Misanin et al. (1968) departed from the usual procedure by replacing the cues for new learning with cues for established memories that they paired with ECS and produced a memory decrement. They also presented ECS without the cue and obtained no memory decrement. Two interpretations of their data are plausible. One is that the ECS reduced the evocative power of the stimulus with which it was paired so that it was not, on a test, followed by the memory, a form of stimulus inhibition; the other is that the memory (response) was blocked by the ECS. The experimenters preferred the second interpretation and concluded that ECS worked only on memories that were active and that memory age was irrelevant. The conclusion is that AMs, or those in transition between inactivity and activity, are open to disruption by amnesic agents, and IMs are not.

It is an interesting speculation that the AM disruption is not limited to amnesic agents. Perhaps AMs are open to decrements due to all forgetting agents (Lewis, 1969). Perhaps IMs are protected from forgetting loss, the loss occurring only with memory activity.

Only a little information relevant to this hypothesis can be found in the traditional animal learning literature because insufficient attention has been devoted to the study of forgetting. Spear (1976), whose own work is an exception, could find no evidence that Pavlov, for example, ever manipulated the retention interval. Students of animal learning traditionally did not use the concept of memory, and the word *forgetting* was almost totally absent from their vocabularies. "Response elimination" was their approximation of this problem, and most of the procedures were those of extinction. If a learned response was repetitively evoked without reinforcement, it would eventually extinguish. This, however, was more likely a form of discrimination learning (discriminating nonreinforced from reinforced conditions) rather than forgetting, which operationally involves a retention interval between the end of learning and the test.

The paucity of animal data is a contrast to that existing for humans. Without going into the wealth of detail existing in the study of human forgetting, an attempt is made to summarize the most important determiners of forgetting in human memories, with the view of determining the possible relevance of this body of information to the kind of AM forgetting conceptualized here.

First are cue and context changes (Tulving, 1972). Memories are formed in a stimulus environment, and these stimuli can later serve to reactive the memories. To the extent that the cues and context change, forgetting will occur, that is, memories will not be evoked. Since cues are always changing, there will be a greater cue change with time, and thus forgetting increases over time. Second, forgetting is also caused by interference, which had traditionally been viewed as a phenomenon of retrieval (Peterson, 1977), even though the interfering material

was learned close to original learning and long before the recall test. Interference basically means that the subject does not distinguish adequately the occasion for remembering X rather than Y. This failure to distinguish is based on the similarity of the learned material to the interpolated material. For example, if a subject first learns an association between A and B and then learns an association between A and C, the two associations will interfere with each other when A is presented at recall. The more similar the items (or lists), the greater will be the interference. Thus, similarity is the keystone of forgetting through interference. It may be that the subject will remember the item but will be confused about the appropriate place to use the item. In the laboratory, this is a failure to make a "list differentiation." Coding serves to increase the distinctiveness of the learned material and thus facilitates retrieval.

In summary, human forgetting is considered to have two primary causes: (a) a change in the stimulus conditions from those that were present at the time of learning. This change can be either external or internal to the organism; and (b) interference due to the learning of similar material.

Can these basic principles of forgetting be integrated with those already discussed and applied to the exit of memories from AM for lower animals? The application is indeed speculative and must be treated with great caution, but Spear (1971, 1976, 1978) has shown that such an application can be fruitful.

First, there are cue changes. An active organism is being constantly bombarded with stimuli, many of which are registered in AM (and which simultaneously have a representation in IM, that is, again, that memories are formed quickly and are permanent). Representations may be held in AM for varying lengths of time. Familiar and expected representations come and go with great rapidity. They already fit into existing memory systems, and little new processing is needed. Perhaps the only new learning that occurs upon the appearance of familiar stimuli is that an old item has occurred again. Unexpected and surprising representations are held for a longer period of time. During this time coding occurs, one of whose purposes is to make the representation retrievable by fitting (association, further learning, coding) it into an existing memory assembly. In a sense this coding makes the representation distinctive, for the representation fits into a particular memory assembly because of its similarities with and differences from other memories. If a representation is not adequately coded, its retrievability will be greatly reduced. This coding (or rehearsal) takes time, during which the organism cannot be attending to other stimuli, or there will be interference and retrievability will be reduced.

A novel stimulus evokes the orienting reflex (Pavlov, 1927) that ordinarily produces a focusing on the new stimulus and therefore protects it from interference during the rehearsal or coding. Another novel stimulus registering during this period that requires coding may preempt the rehearsal process and reduce the retrievability of the first representation (Wagner et al., 1973). The closer the distractor stimulus appears in time to the first stimulus, the greater will be the interference and the less the retrievability. Here, the reference is to the interference between new memories in which inadequate processing is prevented by the preemption of the limited capacity AM. The first stimulus representation should still be capable of retrieval if an appropriate probe is used.

One of the principal points of the present interpretation is that AMs are particularly open to interference. This may be at least one of the reasons that the A-B, A-C paradigm produces interference. The A of the A-C learning component reactivates the A-B memory. When A-B and A-C are both active, the interference occurs. An A-B, X-B sequence would produce little interference because the memories are not concurrently active. Immediately after presentation, paired associates are vulnerable to forgetting, and the loss is geometric. Perhaps forgetting will not occur unless memories are reactivated in a context that produces interference. The time-dependent interference in the Peterson and Peterson

(1959) paradigm suggests this. Gordon and Spear (1973a) have shown an AM interference using the cue-dependent amnesia paradigm. After one memory was well learned, they presented the cue for this memory while simultaneously evoking a competing memory. Considerable forgetting ensued under these circumstances, but there was no forgetting if the established memory was inactive. This is a landmark study in showing that AMs interfere with each other, but that an AM does not interfere with an IM. More recently Gordon (1977a) and Gordon and Feldman (1978) have shown that the reinstatement of a passive avoidance memory will interfere with the retention of a new active avoidance memory.

ECS as Interference

There is a growing body of evidence that ECS blocks AMs regardless of whether they are old or new, but IMs are not disturbed. Generalizing this finding, it has been suggested that AMs may be open to disruption by normal forgetting agents, particularly by similar AMs (Gordon & Spear, 1973a). The similarity between normal forgetting and forgetting due to ECS is difficult to see, but it probably lies in the neural activity that must accompany memory activation. Each AM, whether new or established, must involve a distinctive pattern of neural firing that (a) represents an environmental or other event and (b) will be repeated when the event is repeated. The pattern probably involves a characteristic frequency of firing for single neurons as well as a characteristic assemblage of neurons.

After reviewing a great deal of EEG data, John (1972) concluded that

> the evidence which has been presented indicates that when a specific memory is retrieved, a temporal pattern of electrical activity peculiar to that memory is released in numerous regions of the brain. To that released set of waveshapes corresponds the average firing pattern of ensembles of neurons diffusely distributed throughout these widespread anatomical domains. . . . This suggests that during retrieval of a particular memory, a unique and invariant temporal pattern of coherence occurs in the neural discharges averaged across a spatially distributed and diffuse ensemble of neurons, in which the variable activity of an individual neuron is significant primarily insofar as it contributes to the statistics of the population. (p. 862)

It is assumed that similar memories have similar components of neural firing. It is further assumed that when two similar memories are simultaneously active, there is competition and interference between their firing patterns. That is simply to say, two different patterns involving the same neurons, at least to some degree, cannot be firing simultaneously without interference between the two patterns. *The occurrence of one inhibits or blocks the other.* This is the major assumption. When overlapping neural firing patterns are evoked simultaneously, they interfere with each other and reduce the probability that either will fire again under similar circumstances. We know that similar responses and similar memories compete with each other, and we know that memories and responses are based on neural activities. All that is added here is that memory competition is based on neural competition, and that similarity is an important determiner of the competition. In experimental amnesia, the presentation of the learning stimulus activates a memory and its characteristic pattern of neural firing. While the memory pattern is active, another pattern is created by the ECS, which effectively interferes with or blocks the activated memory pattern.

Admittedly, this is highly speculative and is based on little physiological data. But it is not more speculative or based on less data than a theory which maintains that a learning event establishes a physiological structure that is fragile and open to disruption for a period of time and that amnesic agents serve to destroy this process.

The interference-retrieval hypothesis is far from implausible. The basic assumptions are as follows: (a) An active memory requires neural firing, (b) the pattern of neural firing is characteristic of that memory, (c) similar memories share similar neural firing components, (d) similar, but different, memories and neural firing patterns that are active at the same time interfere with each other,

(e) interference between neural firing patterns results in forgetting, and (f) the interference results not in a memory destruction but a blocking.

Conclusion

A point of view has been presented here in an area of investigation that has many perplexities and complexities. Certainly, there are pockets of data that do not, at the moment, fit tidily into the current conception, but the emphasis on retrieval failure as the source of forgetting in psychobiology has been growing and making increasing inroads on the traditional storage-failure approach. (See Spear's excellent new book, 1978.) Retrieval failure is more consonant with the important theorizing of human cognitive psychology and, at the moment, seems to comprehend a large segment of the psychobiological data, and the distinction between AM and IM affords a cohesive framework for conceptualizing this data. The major lesson for psychobiologists is that they would profit from turning a greater amount of their research efforts to the study of retrieval mechanisms and that they should consider the implications for physiological study of a learning engram that is formed permanently in less than a second.

An interesting point that remains to be considered is the distinction, if any, remaining between a consolidation-storage point of view and a retrieval point of view. Without doubt the distinction is becoming blurred. The consolidation-storage approach has moved from conceptualizing a long-delayed fixing of structure produced by learning (Gold & McGaugh, 1975; McGaugh, 1966) to the position of Bloch (1976), who says that consolidation "would be better named the phase of information processing" (p. 583) or Cherkin (1970), who says that anything occurring soon after acquisition is part of the consolidation process. This conceptualization of consolidation is similar to the present approach and, as Bloch points out, requires brain activities that can be prevented or enhanced.

There nevertheless remain several important distinctions between the two. One of these is empirical and the others are conceptual. The empirical distinction concerns the consolidation-storage approach, which argues that the various brain manipulations producing forgetting do so either by eliminating the physical trace of learning (McGaugh, 1966) or by reducing it below some threshold of expression (Cherkin, 1972).

If reminder stimuli reinstate a memory, as they clearly do, then the amnesia or forgetting cannot have been due to insufficient storage. As has already been reviewed, the storage theorists argue that the reminder stimulus (a) is not effective if amnesia is complete and (b) serves simply as another learning trial if amnesia is not complete. Both of these arguments have been rebutted here and elsewhere (Miller & Springer, 1973, 1974; Spear, 1973; Spear & Parsons, 1976). Reminder has been shown following complete amnesia, and reinstatement has a greater effect than a single additional learning trial. The conclusion drawn here is that ECS and other forms of brain intervention block the retrieval of memory, whether old or new, rather than its storage. This is the interpretation that is most consistent with existing data. On this point differences do remain between a storage and retrieval point of view.

The distinction between the two has largely been lost at the conceptual level when the consolidation-storage approach is viewed as a form of information processing. Information processing here has been treated as a form of coding and elaboration of already stored memories. This elaboration is additional learning and it aids in the retrieval of previous learning. In the current conception, a subject learns A (that FS occurred), then he learns B (that it came through the grid bars), then he learns C and D, and so forth, which are new items of learning that give meaning to the total, including A, and that aid in retrieval. Amnesic agents prevent this further processing, and they also block (interfere with) the later expression of A or of any learning that has occurred up to the point of the amnesic agent as long as that new learning is still in AM. If this is now what consolidation means, then it is to that extent a retrieval theory.

Reference Note

1. Wood, F., & Kinsbourne, M. Paper on the subject of the amnesic syndrome in humans, presented at the meeting of the International Neuropsychology Society, Boston, February 1974.

References

Adams, H. E., Hoblit, P. R., & Sutker, P. B. Electroconvulsive shock brain acetylcholinesterase activity and memory. *Physiology and Behavior,* 1969, *4,* 113–116.

Adams, H. E., & Lewis, D. J. Electroconvulsive shock, retrograde amnesia, and competing responses. *Journal of Comparative and Physiological Psychology,* 1962, *55,* 299–301.

Alpern, H. P., & Kimble, D. P. Retrograde amnesic effects of diethyl ether and bis (tribluoroethyl) ether. *Journal of Comparative and Physiological Psychology,* 1967, *63,* 168–171.

Alpern, H. P., & McGaugh, J. L. Retrograde amnesia as a function of duration of electroshock stimulation. *Journal of Comparative and Physiological Psychology,* 1968, *65,* 265–269.

Atkinson, R. C., & Shiffrin, R. M. Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2). New York: Academic Press, 1968.

Atkinson, R. C., & Shiffrin, R. M. The control of short-term memory. *Scientific American,* 1971, *225*(2), 82–90.

Azmitia, E. C., McEwen, B. S., & Quartermain, D. Prevention of ECS-induced amnesia by reestablishing continuity with the training situation. *Physiology and Behavior,* 1972, *8,* 853–855.

Baddeley, A. D., & Hitch, G. Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 3). New York: Academic Press, 1974.

Barondes, S. H., & Cohen, H. D. Arousal and the conversion of "short-term" to "long-term" memory. *Proceedings of the National Academy of Sciences,* 1968, *61,* 923–929. (a)

Barondes, S. H., & Cohen, H. D. Memory impairment after subcutaneous injection of acetoxycycloheximide. *Science,* 1968, *160,* 556–557. (b)

Barraco, R. A., & Stettner, L. J. Antibiotics and memory. *Psychological Bulletin,* 1976, *83,* 242–302.

Bartus, R. T., & LeVere, T. E. Storage and utilization of information within a discrimination trial. In D. L. Medin, W. A. Roberts, & R. T. Davis (Eds.), *Processes of animal memory.* Hillsdale, N.J.: Erlbaum, 1976.

Bellezza, F. S., & Walker, R. J. Storage-coding tradeoff in short-term store. *Journal of Experimental Psychology,* 1974, *102,* 629–33.

Bjork, R. A. Short-term storage: The ordered output of a central processor. In F. Restle, R. M. Shiffrin, N. J. Castellan, H. R. Lindman, & D. R. Pesoni (Eds.), *Cognitive theory* (Vol. 1). Hillsdale, N.J.: Erlbaum, 1975.

Bloch, V. Brain activation and memory consolidation. In M. R. Rosenzweig & E. L. Bennett (Eds.), *Neural mechanisms of learning and memory.* Cambridge, Mass.: MIT Press, 1976.

Blodgett, H. C. The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology,* 1929, *4,* 113–134.

Blough, E. S. Recognition by the pigeon of stimuli varying in two dimensions. *Journal of the Experimental Analysis of Behavior,* 1972, *18,* 345–367.

Bogoch, S. *The biochemistry of memory.* London: Oxford University Press, 1968.

Bogoch, S. Brain glycoproteins and learning: New studies supporting the "sign-post" theory. In W. B. Essman & S. Nahyima (Eds.), *Current biochemical approaches to learning and memory.* Flushing, N.Y.: Spectrum-Halstead, 1973.

Bower, G. H. Stimulus-sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), *Coding process in human memory.* Washington, D.C.: V. H. Winston, 1972.

Bower, G. H. Cognitive psychology: An introduction. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Introduction to concepts and issues* (Vol. 1). Hillsdale, N.J.: Erlbaum, 1975.

Botwinick, C. Y., & Quartermain, D. Recovery from amnesia induced by pre-test injections of monoamine oxidase inhibitors. *Pharmacology, Biochemistry and Behavior,* 1974, *2,* 375–379.

Braun, J. J., Meyer, P. M., & Meyer, D. R. Sparing of a brightness habit in rats following visual decortication. *Journal of Comparative and Physiological Psychology,* 1966, *61,* 79–82.

Broadbent, D. E. *Perception and communication.* New York: Pergamon Press, 1958.

Brogden, W. J. Sensory pre-conditioning. *Journal of Experimental Psychology,* 1939, *25,* 323–332.

Brown, J. Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology,* 1958, *10,* 12–21.

Buckholtz, N. S., & Bowman, R. E. Incubation and retrograde amnesia studies with various ECS intensities and durations. *Physiology and Behavior,* 1972, *8,* 113–117.

Campbell, B. A., & Jaynes, J. Reinstatement. *Psychological Review,* 1966, *73,* 478–480.

Cherkin, A. Kinetics of memory consolidation: Role of amnestic treatment parameters. *Proceedings of the National Academy of Sciences,* 1969, *63,* 1094–1101.

Cherkin, A. Retrograde amnesia: Impaired memory consolidation or impaired retrieval? *Communications in Behavioral Biology,* 1970, *5,* 183–190.

Cherkin, A. Retrograde amnesia in the chick: Resistance to the reminder effect. *Physiology and Behavior,* 1972, *8,* 949–955.

Chorover, S. An experimental critique of "consolidation studies" and an alternative "model-systems"

approach to the biophysiology of memory. In M. R. Rosenzweig & E. L. Bennett (Eds.), *Neural mechanisms of learning and memory*. Cambridge, Mass.: MIT Press, 1976.

Chorover, S. L., & Schiller, P. H. Short-term retrograde amnesia in rats. *Journal of Comparative and Physiological Psychology*, 1965, *59*, 73–78.

Conrad, R. Acoustic confusions and memory span for words. *Nature*, 1963, *197*, 1029–1030.

Coons, E. E., & Miller, N. E. Conflict versus consolidation of memory traces to explain "retrograde amnesia" produced by ECS. *Journal of Comparative and Physiological Psychology*, 1961, *53*, 524–531.

Corkin, S. Acquisition of motor skills after bilateral medial temporal lobe excision. *Neuropsychologia*, 1968, *6*, 255–265.

Craik, F. I. M. Human memory. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology*. Palo Alto, Calif.: Annual Reviews, 1979.

Craik, F. I. M., & Lockhart, R. S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 1972, *11*, 671–684.

Craik, F. I. M., & Watkins, M. J. The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 599–607.

Davis, R. E., & Hirtzel, M. S. Environmental control of ECS-produced retrograde amnesia in goldfish. *Physiology and Behavior*, 1970, *5*, 1089–1092.

Davis, R. E., & Klinger, P. D. Environmental control of amnesic effects of various agents in goldfish. *Physiology and Behavior*, 1969, *4*, 269–271.

Dawson, R. G., & McGaugh, J. L. Electroconvulsive shock effects on a reactivated memory trace: Further examination, *Science*, 1969, *166*, 525–527. (a)

Dawson, R. G., & McGaugh, J. L. Electroconvulsive shock-produced retrograde amnesia: Analysis of the familiarization effect. *Communications in Behavioral Biology*, 1969, 91–95 (b).

Deutsch, D. Tones and numbers: Specificity of interference in immediate memory. *Science*, 1970, *168*, 1604–1605.

DeVietti, T. L., Conger, G. L., & Kirkpatrick, B. R. Comparison of the enhancement gradients of retention obtained with stimulation of the mesencephalic reticular formation after training or memory reactivation. *Physiology and Behavior*, 1977, *19*, 549–554.

DeVietti, T. L., & Holliday, J. H. Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace: A replication. *Psychonomic Science*, 1972, *29*, 137–138.

DeVietti, T. L., & Kirkpatrick, B. R. The amnesia gradient: Inadequate as evidence for a memory consolidation process. *Science*, 1976, *194*, 438–439.

Duncan, C. P. The retroactive effects of shock on learning. *Journal of Comparative and Physiological Psychology*, 1949, *42*, 32–34.

Ebbinghaus, H. *Memory: A contribution to experimental psychology*. New York: Columbia University Teacher's College, Bureau of Publications, 1913.

Estes, W. K. Toward a statistical theory of learning. *Psychological Review*, 1950, *57*, 94–107.

Everett, J. C., & Corson, J. A. ECS in one-trial appetitive learning in rats. *Journal of Comparative and Physiological Psychology*, 1973, *84*, 353–360.

Flexner, J. B., Flexner, L. B., & Stellar, E. Memory in mice as affected by intracerebral puromycin. *Science*, 1963, *141*, 57–59.

Flexner, L. B., Flexner, J. B., & Roberts, R. B. Memory in mice analyzed with antibiotics. *Science*, 1967, *155*, 1377–1383.

Garcia, J., & Koelling, R. A. Relation of cue to consequences in avoidance learning. *Psychonomic Science*, 1966, *4*, 123–124.

Garcia, J., Kovner, R., & Green, K. F. Cue properties vs. palatability of flavours in avoidance learning. *Psychonomic Science*, 1970, *20*, 313–319.

Garcia, J., McGowan, B. K., Erwin, F. R., & Koeuer, R. A. Cues: Their effectiveness as a function of the reinforcer. *Science*, 1968, *160*, 794–795.

Gardner, H., Boller, F., Moreines, J., & Butters, N. Retrieving information from Korsakoff patients: Effects of categorized cues and references to the task. *Cortex*, 1973, *9*, 165–175.

Gardner, W. R. The stimulus in information processing. *Amercian Psychologist*, 1970, *25*, 350–358.

Geller, A., & Jarvik, M. E. The time relations of ECS-induced amnesia. *Psychonomic Science*, 1968, *12*, 169–170.

Glanzer, M. Storage mechanisms in recall. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 5). New York: Academic Press, 1972.

Glickman, S. Deficits in avoidance learning produced by stimulation of the ascending reticular formation. *Canadian Journal of Psychology*, 1958, *12*, 97–102.

Glickman, S. Perseverative neural responses and consolidation of the neural trace. *Psychological Bulletin*, 1961, *58*, 218–233.

Gold, P. E., & McGaugh, J. L. A single-trace, two-process view of memory storage process. In D. Deutsch & J. A. Deutsch (Eds.), *Short-term memory*. New York: Academic Press, 1975.

Gordon, W. C. Similarities between recently acquired and reactivated memories with production of memory interference. *American Journal of Psychology*, 1977, *90*, 231–242. (a)

Gordon, W. C. Susceptibility of a reactivated memory to the effects of strychnine: A time-dependent phenomenon. *Physiology and Behavior*, 1977, *18*, 95–99. (b)

Gordon, W. C., & Feldman, D. T. Reaction induced interference in a short-term retention paradigm. *Learning and Motivation*, 1978, *9*, 164–178.

Gordon, W. C., & Spear, N. E. Effect of reactivation of a previously acquired memory on the interaction between memories in the rat. *Journal of Experimental Psychology*, 1973, *99*, 349–355. (a)

Gordon, W. C., & Spear, N. E. The effect of strychnine on recently acquired and reactivated passive

avoidance memories. *Physiology and Behavior,* 1973, *10,* 1071–1075. (b)

Greeno, J. G. How associations are memorized. In D. A. Norman (Ed.), *Models of human memory.* New York: Academic Press, 1970.

Grice, G. R. The relation of secondary reinforcement to delayed reward in visual discrimination learning. *Journal of Experimental Psychology,* 1948, *38,* 1–16.

Guthrie, E. R. *The psychology of learning.* New York: Harper, 1935.

Hearst, E., & Peterson, G. B. Transfer of conditioned excitation and inhibition from one operant response to another. *Journal of Experimental Psychology,* 1973, *99,* 360–368.

Hebb, D. O. *The organization of behavior.* New York: Wiley, 1949.

Hill, W. F., & Spear, N. E. Resistance to extinction as a joint function of reward magnitude and the spacing of extinction trials. *Journal of Experimental Psychology,* 1962, *64,* 636–639.

Hinderliter, C. F., Smith, S. G., & Misanin, J. R. Effects of pretraining experience on retention of a passive avoidance task following ECS. *Physiology and Behavior,* 1973, *10,* 671–675.

Hintzman, D. L. Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium.* Hillsdale, N.J.: Erlbaum, 1974.

Howard, R. L., Glendenning, R. L., & Meyer, D. R. Motivational control of retrograde amnesia: Further explorations and effects. *Journal of Comparative and Physiological Psychology,* 1974, *86,* 187–192.

Howard, R. L., & Meyer, D. R. Motivational control of retrograde amnesia in rats: A replication and extension. *Journal of Comparative and Physiological Psychology,* 1971, *74,* 37–40.

Hudspeth, W. J., McGaugh, J. L., & Thomson, C. W. Aversive and amnesic effects of electroconvulsive shock. *Journal of Comparative and Physiological Psychology,* 1964, *57,* 61–64.

Hull, C. L. *Principles of behavior.* New York: Appleton-Century-Crofts, 1943.

Hunter, W. S. The delayed reaction in animals and children. *Behavior Monographs,* 1913, *2*(2, Serial No. 6).

Irwin, S., Banuazizi, A., Kalsner, S., & Curtis, A. One-trial learning in the mouse: Its characteristics and modifications by experimental-seasonal variables. In A. G. Karczmay & W. P. Koella (Eds.), *Neurophysiological and behavioral aspects of psychotropic drugs.* Springfield, Ill.: Charles C Thomas, 1969.

Isaacson, R. L., & Pribram, K. H. *The hippocampus* (Vol. 2). New York: Plenum Press, 1975.

James, W. *The principles of psychology.* New York: Holt, 1890.

Jensen, R. A., & Riccio, D. Effects of prior experience upon retrograde amnesia produced by hypothermia. *Physiology and Behavior,* 1970, *5,* 1291–1294.

John, E. R. Switchboard versus statistical theories of learning and memory. *Science,* 1972, *177,* 850–864.

Keith-Lucas, T., & Guttman, N. Robust single-trial delayed backward conditioning. *Journal of Comparative and Physiological Psychology,* 1975, *88,* 468–496.

Keppel, G., & Underwood, B. J. Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior,* 1962, *1,* 153–161.

Kesner, R. A neural system analysis of memory storage and retrieval. *Psychological Bulletin,* 1973, *80,* 177–203.

Kesner, R. P., & Connor, H. S. Independence of short- and long-term memory: A neural systems approach. *Science,* 1972, *176,* 432–434.

Kesner, R. P., & Connor, H. S. Effects of electrical stimulation of rat limbic system and midbrain reticular formation upon short- and long-term memory. *Physiology and Behavior,* 1974, *12,* 5–12.

Kety, S. S. The biogenic amines in the central nervous system: Their possible roles in arousal, emotion, and learning. In F. O. Schmitt (Ed.), *The neurosciences: Second study program.* New York: Rockefeller University Press, 1970.

Kinsbourne, M., & Wood, F. Short-term memory processes and the amnesic syndrome. In D. Deutsch & J. A. Deutsch (Eds.), *Short-term memory.* New York: Academic Press, 1975.

Kopp, R., Bohdanecky, Z., & Jarvik, M. E. Long temporal gradients of retrograde amnesia for a well-discriminated stimulus. *Science,* 1966, *153,* 1547–1549.

Koppenaal, R. J., Jagoda, E. R., & Cruce, J. A. F. Recovery from ECS-produced amnesia following a reminder. *Psychonomic Science,* 1967, *9,* 293–294.

Landauer, T. K. Consolidation in human memory: Retrograde amnestic effects of confusable items in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior,* 1974, *13,* 45–53.

Landfield, P. W., McGaugh, J. L., & Tusa, R. J. Theta rhythm: A temporal correlate of memory storage processes in the rat. *Science,* 1972, *175,* 87–89.

Leith, C. R., & Mahi, W. S., Jr. Effects of compound configuration on stimulus selection in the pigeon. *Journal of Experimental Psychology: Animal Behavior Processes,* 1977, *3,* 229–239.

Lett, B. T. Delayed reward learning: Disproof of the traditional theory. *Learning and Motivation,* 1973, *4,* 237–246.

Lett, B. T. Long delay learning: Implications for learning and memory theory. In N. S. Sutherland (Ed.), *Tutorial essays in experimental psychology.* New York: Academic Press, 1978.

Lewis, D. J. Sources of experimental amnesia. *Psychological Review,* 1969, *76,* 461–472.

Lewis, D. J. A cognitive approach to experimental amnesia. *American Journal of Psychology,* 1976, *89,* 51–80.

Lewis, D. J., & Bregman, N. The source of the cues for cue-dependent amnesia. *Journal of Comparative and Physiological Psychology,* 1973, *85,* 421–426.

Lewis, D. J., Bregman, N. J., & Mahan, J. J., Jr. Cue-dependent amnesia in rats. *Journal of Comparative and Physiological Psychology*, 1972, *81*, 243–247.

Lewis, D. J., & Maher, B. A. Neural consolidation and electroconvulsive shock. *Psychological Review*, 1965, *72*, 225–239.

Lewis, D. J., & Maher, B. A. Electroconvulsive shock and inhibition: Some problems considered. *Psychological Review*, 1966, *73*, 388–392.

Lewis, D. J., Miller, R. R., & Misanin, J. R. Control of retrograde amnesia. *Journal of Comparative and Physiological Psychology*, 1968, *66*, 48–52.

Lewis, D. J., Miller, R. R., & Misanin, J. R. Selective amnesia in rats produced by electroconvulsive shock. *Journal of Comparative and Physiological Psychology*, 1969, *69*, 136–140.

Lewis, D. J., & Nicholas, T. Amnesia for active memory. *Physiology and Behavior*, 1973, *11*, 821–825.

Lubow, R. E., Rifkin, B., & Alek, M. The context effect: The relationship between stimulus preexposure and environmental preexposure determines subsequent learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 1976, *2*, 38–47.

Luttges, M. W., & McGaugh, J. L. Permanence of retrograde amnesia produced by electroconvulsive shock. *Science*, 1967, *156*, 408–410.

Lynch, G., Deadwyler, S., & Cotman, C. Postlesion axonal growth produces permanent functional connections. *Science*, 1973, *180*, 1364–1366.

Madigan, S. A., & McCabe, L. Perfect recall and total forgetting: A problem for models of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 1971, *10*, 101–106.

Mah, C. S., & Albert, D. J. Electroconvulsive shock-induced amnesia gradient. *Behavioral Biology*, 1973, *9*, 517–540.

Mahut, H. Effects of subcortical electrical stimulation on learning in the rat. *Journal of Comparative and Physiological Psychology*, 1962, *55*, 472–477.

McDonough, J. H., & Kesner, R. P. Amnesia produced by brief electrical stimulation of the amygdala or dorsal hippocampus in cats. *Journal of Comparative and Physiological Psychology*, 1971, *77*, 171–178.

McGaugh, J. L. Time dependent processes in memory storage. *Science*, 1966, *153*, 1351–1358.

McGaugh, J. L., & Dawson, R. G. Modification of memory storage processes. *Behavioral Sciences*, 1971, *16*, 45–63.

McGaugh, J. L., & Gold, P. E. Modulation of memory by electrical stimulation of the brain. In M. R. Rosenzweig & E. L. Bennett (Eds.), *Neural mechanisms in learning and memory*. Cambridge, Mass.: MIT Press, 1976.

Melton, A. W. Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 1963, *9*, 596–606.

Meyer, D. R., & Beattie, M. S. Some properties of substrates of memory. In L. Miller, C. Sandman, & A. Kasten (Eds.), *Neuropeptide influences on brain and behavior*. New York: Academic Press, 1977.

Miller, G. A. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 1956, *63*, 81–96.

Miller, R. R. Effects of environmental complexity on amnesia induced by electroconvulsive shock in rats. *Journal of Comparative and Physiological Psychology*, 1970, *71*, 267–275.

Miller, R. R., Misanin, J. R., & Lewis, D. J. Amnesia as a function of events during the learning-ECS interval. *Journal of Comparative and Physiological Psychology*, 1969, *67*, 145–148.

Miller, R. R., & Springer, A. D. Induced recovery of memory in rats following electroconvulsive shock. *Physiology and Behavior*, 1972, *8*, 645–651.

Miller, R. R., & Springer, A. D. Amnesia, consolidation, and retrieval. *Psychological Review*, 1973, 80, 69–79.

Miller, R. R., & Springer, A. D. Implications of recovery from experimental amnesia. *Psychological Review*, 1974, *81*, 470–473.

Milner, B. Amnesia following operation on the temporal lobes. In C. W. M. Whitty & O. L. Zangwill (Eds.), *Amnesia*. London: Butterworths, 1966.

Milner, B. Memory and the medial temporal regions of the brain. In K. Pribram & D. Broadbent (Eds.), *Biology of memory*. New York: Academic Press, 1970.

Misanin, J. R., Miller, R. R., & Lewis, D. J. Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace. *Science*, 1968, *160*, 554–555.

Mitchell, D., Scott, D. W., & Mitchell, L. K. Attenuated and enhanced neophobia in the taste-aversion "delay of reinforcement" effect. *Animal Learning and Behavior*, 1977, *5*, 99–102.

Müller, G. E., & Pilzecker, A. Experimental beiträge zue lehre bom gedäctnesses. *Zeitschrift für psychologie*, 1900, *1*, 1–288.

Murdock, B. B., Jr. *Human memory: Theory and data*. Hillsdale, N.J.: Erlbaum, 1974.

Nicholas, T., Galbraith, G., & Lewis, D. J. Theta activity and memory processes in rats. *Physiology and Behavior*, 1976, *16*, 489–492.

Paolino, R. M., Quartermain, D., & Miller, N. E. Different temporal gradients of retrograde amnesia produced by carbon dioxide anesthesia and electroconvulsive shock. *Journal of Comparative and Physiological Psychology*, 1966, *62*, 270–274.

Pavlov, I. P. *Conditioned reflexes*. London: Oxford University Press, 1927.

Pearlman, C. A., Sharpless, S. K., & Jarvik, M. E. Retrograde amnesia produced by anesthetic and convulsant agents. *Journal of Comparative and Physiological Psychology*, 1961, *54*, 109–112.

Peterson, L. R. Verbal learning and memory. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology*. Palo Alto, Calif.: Annual Reviews, 1977.

Peterson, L. R., & Peterson, M. J. Short-term retention of individual verbal terms. *Journal of Experimental Psychology*, 1959, *58*, 193–198.

Posner, M. I. Short-term memory systems in human

information processing. In A. F. Sanders (Ed.), *Attention and performance* (Vol. 1). Amsterdam: North-Holland, 1967.

Posner, M. I. *Cognition: An introduction*. Glenview, Ill.: Scott, Foresman, 1973.

Potts, W. J. The effect of different environments prior to electroconvulsive shock on the gradient of retrograde amnesia. *Physiology and Behavior*, 1971, *7*, 61–164.

Quartermain, D. The influence of drugs on learning and memory. In M. R. Rosenzweig & E. L. Bennett (Eds.), *Neural mechanism of learning and memory*. Cambridge, Mass.: MIT Press, 1976.

Quartermain, D., McEwen, B. S., & Azmitia, E. C., Jr. Amnesia produced by electroconvulsive shock or cycloheximide: Conditions for recovery. *Science*, 1970, *169*, 683–686.

Reitman, J. S. Without surreptitious rehearsal, information in short-term memory decays. *Journal of Verbal Learning and Verbal Behavior*, 1974, *13*, 365–377.

Riccio, D. C., & Stikes, E. R. Persistent but modifiable retrograde amnesia produced by hypothermia. *Physiology and Behavior*, 1969, *4*, 649–652.

Rigter, H., & Van Riezen, H. Anti-amnesic effect of $ACTH_{4-10}$: Its independence of the nature of the amnesic agent and the behavioral test. *Physiology and Behavior*, 1975, *14*, 563–566.

Robbins, M. J., & Meyer, D. R. Motivational control of retrograde amnesia. *Journal of Experimental Psychology*, 1970, *84*, 220–225.

Roberts, R. B., Flexner, J. B., & Flexner, L. B. Some evidence for the involvement of adrenergic sites in the memory trace. *Proceedings of the National Academy of Sciences*, 1970, *66*, 310–313.

Routtenberg, A., & Holtzman, N. Memory disruption by electrical stimulation of substantia nigra, pars compacta. *Science*, 1973, *181*, 83–86.

Rutledge, L. T. Synaptogenesis: Effects of synaptic use. In M. R. Rosenzweig & E. L. Bennett (Eds.), *Neural mechanisms of learning and memory*. Cambridge, Mass.: MIT Press, 1976.

Sara, S. J., David-Remacle, M., & Lefevre, D. Passive avoidance behavior in rats after electroconvulsive shock: Facilitative effect of response retardation. *Journal of Comparative and Physiological Psychology*, 1975, *89*, 489–497.

Schneider, A. M., & Sherman, W. Amnesia: A function of the temporal relation of footshock to electroconvulsive shock. *Science*, 1968, *159*, 219–221.

Seligman, M. E. P. On the generality of the laws of learning. *Psychological Review*, 1970, *77*, 406–418.

Serota, R. G., Roberts, R. B., & Flexner, L. B. Acetoxycycloheximide-induced transient amnesia: Protective effects of adrenergic stimulants. *Proceedings of the National Academy of Sciences*, 1972, *69*, 340–342.

Shiffrin, R. M. Capacity limitations in information processing, attention, and memory. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Vol. 4. Attention and memory*. Hillsdale, N.J.: Erlbaum, 1976.

Shulman, H. G. Encoding and retention of semantic and phonemic information in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 1970, *9*, 499–508.

Shulman, H. G. Semantic confusion errors in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 1972, *11*, 221–227.

Solomon, R. L., & Turner, L. H. Discriminative classical conditioning in dogs paralyzed by curare can later control discriminative avoidance responses in the normal state. *Psychological Review*, 1962, *69*, 202–219.

Spear, N. E. Forgetting as retrieval failure. In W. K. Honig & P. H. R. James (Eds.), *Animal memory*. New York: Academic Press, 1971.

Spear, N. E. Retrieval of memory in animals. *Psychological Review*, 1973, *80*, 163–194.

Spear, N. E. Retrieval of memories: A psychobiological approach. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes*. Hillsdale, N.J.: Erlbaum, 1976.

Spear, N. E. *The processing of memories: Forgetting and retention*. Hillsdale, N.J.: Erlbaum, 1978.

Spear, N. E., & Parsons, P. J. Analysis of a reactivation treatment: Ontogenetic determinants of alleviated forgetting. In D. L. Medin, W. A. Roberts, & R. T. Davis (Eds.), *Processes of animal memory*. New York. Wiley, 1976.

Spence, K. W. The role of secondary reinforcement in delayed reward learning. *Psychological Review*, 1947, *54*, 1–8.

Sperber, R. D., Greenfield, D. R., & House, B. J. A nonmonotonic effect of distribution of trials in retardate learning and memory. *Journal of Experimental Psychology*, 1973, *99*, 186–198.

Sperling, G. A model for visual memory tasks. *Human Factors*, 1963, *5*, 19–31.

Squire, L. R. Short-term memory as a biological entity. In D. Deutsch & J. A. Deutsch (Eds.), *Short-term memory*. New York: Academic Press, 1975.

Sternberg, S. High-speed scanning in human memory. *Science*, 1966, *153*, 652–654.

Talland, G. *Disorders of memory and learning*. Baltimore: Penguin Books, 1968.

Terry, W. S., & Wagner, A. R. Short-term memory for "surprising" vs. "expected" US in Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavioral Processes*, 1975, *104*, 122–133.

Thistlethwaite, D. A critical review of latent learning and related experiments. *Psychological Bulletin*, 1951, *48*, 97–129.

Tolman, E. C. *Purposive behavior in rats and men*. New York: Century, 1932.

Tolman, E. C. There is more than one kind of learning. *Psychological Review*, 1949, *56*, 144–155.

Tulving, E. Theoretical issues in free recall. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory*. Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Tulving, E. Short- and long-term memory: Different retrieval mechanisms. In K. H. Pribram & D. E. Broadbent (Eds.), *Biology of memory*. New York: Academic Press, 1970.

Tulving, E. Episodic and semantic memory. In

Tulving & W. Donaldson (Eds.), *Organization of memory*. New York: Academic Press, 1972.

Tulving, E. Cue-dependent forgetting. *American Scientist*, 1974, *62*, 74–82.

Underwood, B. J. Attributes of memory. *Psychological Review*, 1969, *76*, 559–573.

Wagner, A., Rudy, J., & Whitlow, J. Rehearsal in animal conditioning. *Journal of Experimental Psychology*, 1973, *97*, 407–426.

Warrington, E., & Weiskrantz, L. Verbal learning and retention by amnesic patients using partial information. *Psychonomic Science*, 1970, *20*, 210–211.

Waugh, N. C., & Norman, D. A. Primary memory. *Psychological Review*, 1965, *72*, 89–104.

Weiskrantz, L., & Warrington, E. K. The problem of the amnesic syndrome in man and animals. In R. L. Isaacson & K. H. Pribram (Eds.), *The hippocampus* (Vol. 2). New York: Plenum Press, 1975.

Wickelgren, W. A. Similarity and intrusions in short-term memory for consonant–vowel trigrams *Quarterly Journal of Experimental Psychology*, 1965, *17*, 241–246.

Wickelgren, W. A., & Norman, D. A. Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology*, 1966, *3*, 316–347.

Wilburn, M. W., & Kesner, R. P. Differential amnesia effects produced by electrical stimulation of the caudate nucleus and nonspecific thalamic systems. *Experimental Neurology*, 1972, *34*, 45–50.

Received May 5, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1084–1089

Interactions, Partial Interactions, and Interaction Contrasts in the Analysis of Variance

Robert J. Boik
Pepperdine University

A two-step framework for the interpretation of significant two-treatment interactions is proposed. First, a contrast between the means of one treatment is estimated separately at each level of the second treatment. A partial interaction test tests the hypothesis that the variabiliy among these contrasts is zero. The second step consists of computing a difference between the separately estimated contrasts. An interaction contrast test tests the hypothesis that this difference is zero. The familywise Type I error rate can be controlled at alpha by employing Gabriel's simultaneous test procedure for partial interaction tests and Scheffé's method for interaction contrast tests.

Recently, an issue has arisen concerning a posteriori tests after detection of a significant interaction in the analysis of variance. Marascuilo and Levin (1970) and Levin and Marascuilo (1972) have criticized the use of simple effects tests in which the means within a single row or column of the data matrix are compared. Their basic criticism is that the null hypotheses tested by simple effects tests are not coherent with the null hypothesis tested by the a priori omnibus interaction test. To maintain a coherent analysis, Marascuilo and Levin recommended testing interaction contrasts after detection of a significant interaction.

In commenting on the approach of Marascuilo and Levin (1970), Games (1973) argued that although interaction contrasts are coherent with the omnibus interaction test, these contrasts do not easily lend themselves to meaningful behavioral interpretations. Consequently, Games recommended performing simple effects tests. Levin and Marascuilo (1973) replied by emphasizing the flexibility of the interaction contrast approach. Thus far, there is little evidence that the issue has been resolved (Marascuilo & Levin, 1976; Games Note 1).

This article neither attempts to resolve nor necessarily even to clarify the issue. The question of whether interaction contrasts should be tested must ultimately be answered by each individual researcher. Rather, the article attempts to present a conceptual framework for testing interaction contrasts, should the researcher decide to do so. Interaction contrasts will not be tested unless they can be readily interpreted by the researcher. It is hoped that this the conceptual framework presented in this article will increase the meaningfulness of interaction contrasts.

Hypothetical Data

The means for a set of hypothetical data are presented in Table 1. Each of 72 biology students complaining of a severe fear of blood (hemophobia) was randomly assigned to one of the 12 cells with the restriction that 6 students be assigned per cell. Treatment A represents three types of fear reduction therapy. Subjects in the first level of Treatment A served as the control group and did not participate in any therapy. Subjects in the second and third levels of Treatment A participated in therapies denoted as a_2 and respectively. Treatment B represents four doses of an antianxiety medication. The dose level increases from b_1 (placebo) to b_4 (ma

A version of this article was presented at the meeting of the Western Psychological Association, San Francisco, April 1978.

Requests for reprints should be sent to Robert J. Boik, Department of Applied Statistics, School of Business, Temple University, Philadelphia, Pennsylvania 19122.

imum dose). Therapy sessions occurred while the subjects were under the influence of the drug. The dependent variable is the magnitude of the electrodermal response (in arbitrary units) when subjects were exposed to blood. Subjects were not under the influence of the drug when the electrodermal response was measured.

The ANOVA summary table is presented in Table 2. Both treatments and the interaction are significant at the .01 level. The strength of the treatments and interaction as measured by ω^2 (see Kirk, 1968, for a description) is as follows: $\omega^2{}_A = .33$, $\omega^2{}_B = .31$, and $\omega^2{}_{AB} = .17$. The strength of association for the treatments and interaction combined is $\omega^2 = .81$. Thus the therapies used, the drug doses selected, and the interaction between the two treatments do account for a substantial portion of the variability among the electrodermal response scores (however, see Glass & Hakstian, 1969, for a cautionary note on the use of ω^2).

Treatment Contrasts

A contrast among the p levels of Treatment A is denoted by A_t and defined as

$$A_t = \sum_{1}^{p} c_i \mu_{i.}, \tag{1a}$$

where t = a subscript of arbitrary value used to distinguish one Treatment A contrast from another; c_i = the coefficient of the contrast associated with the ith level of Treatment A. The p coefficients are subject to the restriction

$$\sum_{1}^{p} c_i = 0. ,$$

Table 1
Hypothetical Means for a 3 × 4 Factorial Design in Which Each Cell Mean Is Based on Six Observations

Treatment A	Treatment B b_1	b_2	b_3	b_4	Row M
a_1	50.2	47.5	46.0	47.9	47.9
a_2	49.9	38.2	28.5	19.0	33.9
a_3	45.7	39.1	36.5	32.7	38.5
Column M	48.6	41.6	37.0	33.2	40.1

Table 2
Summary of Analysis of Variance

Source	*SS*	*df*	*MS*	*F*
Treatment A	2,444.16	2	1,222.08	63.82
Treatment B	2,370.96	3	970.32	41.27
Treatment A × Treatment B	1,376.40	6	229.50	11.98
Within cell	1,149.00	60	19.15	

Note. Total *SS*s = 7,340.52, total *df*s = 71. All *F*s are significant at $p \leq .01$.

and $\mu_{i.}$ = the population mean of the ith level of Treatment A averaged across all levels of Treatment B. The estimate of A_t is denoted by $\hat{A}_t$ and defined as

$$\hat{A}_t = \sum_{1}^{p} c_i M_{i.}\ , \tag{1b}$$

where $M_{i.}$ = the sample mean of the ith level of Treatment A averaged across all levels of Treatment B. The $p \times 1$ vector of coefficients (i.e., the p c_is) associated with A_t and $\hat{A}_t$ is denoted by $\mathbf{A}_t$. For example, the coefficient vectors associated with the three pairwise Treatment A contrasts are $\mathbf{A}'_1 = (1\ -1\ 0)$, $\mathbf{A}'_2 = (1\ 0\ -1)$, and $\mathbf{A}'_3 = (0\ 1\ -1)$. From Equation 1b, the estimates of the pairwise contrasts are $\hat{A}_1 = 14.0$, $\hat{A}_2 = 9.4$, and $\hat{A}_3 = -4.6$.

The sum of squares accounted for by a Treatment A contrast is given by

$$SS_{A_t} = \frac{nq\hat{A}_t^2}{\sum_{1}^{p} c_i^2}, \tag{2}$$

where q is equal to the number of levels of B. From Equation 2, the sums of squares of the three pairwise contrasts are $SS_{A_1} = 2352.00$, $SS_{A_2} = 1060.32$, and $SS_{A_3} = 253.92$. The corresponding 1 and 60 degrees of freedom F ratios are $F_{A_1} = 2352.00/19.15 = 122.82$, $F_{A_2} = 1060.32/19.15 = 55.37$, and $F_{A_3} = 253.92/19.15 = 13.26$. The critical F using Scheffé's (1953) method and $\alpha = .01$ is $S = 2 \cdot F(2, 60) = 2(4.98) = 9.96$. It is concluded that both therapies (a_2 and a_3) resulted in a smaller electrodermal response (EDR) than the control group and that therapy a_2 resulted in a smaller EDR than therapy a_3.

A contrast between the q levels of Treatment

B is denoted by B_u and defined as

$$B_u = \sum_1^q c_j \mu_{\cdot j}\,, \tag{3a}$$

where u = a subscript of arbitrary value used to distinguish one Treatment B contrast from another, and c_j = the coefficient of the contrast associated with the jth level of Treatment B. The q coefficients are subject to the restriction

$$\sum_1^q c_j = 0.$$

$\mu_{\cdot j}$ = the population mean of the jth level of Treatment B averaged across all levels of Treatment A. The estimate of B_u is denoted by $\hat{B}_u$ and defined as

$$\hat{B}_u = \sum_1^q c_j M_{\cdot j}\,, \tag{3b}$$

where $M_{\cdot j}$ = the sample mean of the jth level of Treatment B averaged across all levels of Treatment A. The $q \times 1$ vector of coefficients associated with B_u and $\hat{B}_u$ is denoted by $\mathbf{B}_u$. For example, the coefficient vector associated with the contrast between the placebo dose (b_1) condition and the three nonzero dose conditions is $\mathbf{B}'_1 = (3\ -1\ -1\ -1)$. The contrast estimate from Equation 3b, is $\hat{B}_1 = 34.0$.

The sum of squares accounted for by a Treatment B contrast is given by

$$SS_{B_u} = \frac{np\hat{B}_u^2}{\sum_1^q c_j^2}\,. \tag{4}$$

For example, the sum of squares of the B_1 contrast is $SS_{B_1} = 1734.00$. The F ratio is $F_{B_1} = 1724.00/19.5 = 90.55$. Since the observed F is larger than the Scheffé value, $S = 3F(3, 60) = 3(4.13) = 12.39$, $\alpha = .01$, it is concluded that on the average, the EDR is smaller under the three drug conditions than under the placebo condition.

Partial Interactions and Interaction Contrasts

Often treatment contrasts such as A_1, A_2, A_3, and B_1 are not tested when the Treatment A $\times$ Treatment B interaction is significant. This reflects the commonly held belief that a significant interaction renders treatment contrasts meaningless. Fortunately, this is not the case.

Let us define $A_{t(j)}$ as a simple Treatment A contrast at the jth level of Treatment B. That is,

$$A_{t(j)} = \sum_1^p c_i \mu_{ij}\,, \tag{5a}$$

where μ_{ij} = the population mean of the ith level of Treatment A at the jth level of Treatment B. The simple treatment contrast $A_{t(j)}$ is estimated by $\hat{A}_{t(j)}$, where

$$\hat{A}_{t(j)} = \sum_1^p c_i M_{ij}\,, \tag{5b}$$

and M_{ij} = the sample mean of the ith level of Treatment A at the jth level of Treatment B. For example, from Equation 5b, the estimates of the $A_{2(j)}$ simple treatment contrast are

$$\hat{A}_{2(1)} = 4.50,$$
$$\hat{A}_{2(2)} = 8.40,$$
$$\hat{A}_{2(3)} = 9.5,$$

and

$$\hat{A}_{2(4)} = 15.2.$$

The simple effects procedure described by Games (1973) consists of individually testing the four simple treatment contrasts while ignoring the A_2 treatment contrast itself. However, the significant Treatment A $\times$ Treatment B interaction does not necessarily indicate heterogeneity among the q $A_{2(j)}$ simple contrasts. If the observed variability among the four $\hat{A}_{2(j)}$ simple contrast estimates can be attributed to experimental error, then the A_2 treatment contrast can be interpreted without regard for the significant Treatment A $\times$ Treatment B interaction. The test of homogeneity among the four $A_{2(j)}$ simple contrasts represents a partial test of this interaction. Let us denote this source of variation as the interaction between the A_2 contrast and Treatment B or, more simply the A_2B partial interaction.

The sum of squares for an A_tB partial interaction can be obtained from

$$SS_{A_t\mathrm{B}} = n\{\sum_1^q \hat{A}^2_{t(j)} - [\sum_1^q \hat{A}_{t(j)}]^2/q\}\Big/\sum_1^p c_i^2\,.$$

Equation 6 represents a modification of the familiar computational formula for a sum of squares:

$$SS = \sum_1^n X_m^2 - (\sum_1^n X_m)^2/n.$$

The simple contrast estimates have been substituted for the raw scores; the equation has been divided by

$$\sum_1^p c_i^2$$

to normalize the contrasts; and finally, because each cell mean is based on n observations, the equation has been multiplied by n. The degrees of freedom for the A_tB partial interaction are obtained by multiplying the degrees of freedom of the A_t contrast by the degrees of freedom for Treatment B. That is,

$$df_{A_t\mathrm{B}} = (df_{A_t}) \cdot (df_{\mathrm{B}}). \qquad (7)$$

From Equation 6, the sum of squares of the A_2B partial interaction is $SS_{A_2\mathrm{B}} = 6(412.1 - 37.6^2/4)/2 = 175.98$. The degrees of freedom from Equation 7 are $df_{A_2\mathrm{B}} = (1)(3) = 3$. The mean square and F ratio are therefore $MS_{A_2\mathrm{B}} = 175.98/3 = 58.66$, and $F_{A_2\mathrm{B}} = 58.66/19.15 = 3.06$.

The critical value for the F ratio of an a posteriori partial interaction test cannot be obtained by Scheffé's method unless the partial interaction has only one degree of freedom. However, the simultaneous test procedure (STP) developed by Gabriel (1964, 1969) can be employed. The STP and Scheffé's method are both coherent with the omnibus F test and therefore are coherent with each other. The relationship between Scheffé's method and the STP has been described by Boik (1979). The critical value for an a posteriori ν_3 degrees of freedom partial interaction test is

$$G_{\nu_3} = \nu_1 F(\nu_1, \nu_2)/\nu_3, \qquad (8)$$

where $F(\nu_1, \nu_2)$ = the critical value for the omnibus interaction test. Note that when $\nu_3 = \nu_1$, the critical value is equivalent to that of the omnibus test, and when $\nu_3 = 1$, the critical value is equivalent to that given by Scheffé's method.

For the A_2B partial interaction test, the critical STP value is $G_3 = [6 \cdot F(6, 60)]/3 = (6 \cdot 3.12)/3 = 6.24$, $\alpha = .01$. Since the observed F of 3.06 is less than 6.24, it is concluded that the $A_{2(j)}$ simple contrasts are homogeneous over the q levels of Treatment B. Thus, not only does the a_2 therapy result in a lower EDR than the control group on the average, but in addition, the difference is the same for all four drug levels.

A similar analysis for the A_1 contrast results in a different conclusion. The sample estimates of the four $A_{1(j)}$ simple treatment contrasts are

$$\hat{A}_{1(1)} = .3,$$
$$\hat{A}_{1(2)} = 9.3,$$
$$\hat{A}_{1(3)} = 17.5,$$

and

$$\hat{A}_{1(4)} = 28.9.$$

From Equations 6 and 7, the sum of squares and degrees of freedom for the A_1B partial interaction are $SS_{A_1\mathrm{B}} = 1332.12$, and $df_{A_1\mathrm{B}} = 3$. The mean square and F are therefore $MS_{A_1\mathrm{B}} = 444.04$, and $F_{A_1\mathrm{B}} = 23.19$. Since the observed F is larger than the critical STP value of 6.24, the A_1B partial interaction is judged significant. This indicates that the difference between the control condition and therapy a_2 is not identical for all drug levels. An examination of the $A_{1(j)}$ simple contrast estimates suggests that the difference between the control group and the a_2 therapy group might be larger under the three nonzero drug dose conditions than under the placebo condition. This hypothesis can be tested by means of an interaction contrast.

A product interaction contrast consists of a contrast between simple treatment contrasts. In the previous example, the interaction contrast described consists of the difference between the $A_{1(j)}$ simple contrast under the placebo condition and the average $A_{1(j)}$ simple contrast under the three drug conditions. The Treatment B coefficient vector associated with the interaction contrast is $\mathbf{B}'_1 = (3\ -1\ -1\ -1)$.

Let us denote a product interaction contrast as $A_t B_u$ and define it as

$$A_t B_u = \sum_1^q c_j A_{t(j)}\,, \qquad (9a)$$

where c_j = the jth coefficient in the coefficient vector $\mathbf{B}_u$. The coefficient vector $\mathbf{B}_u$ describes

the contrast between the $A_{t(j)}$ simple treatment contrasts. An A_tB_u interaction contrast is estimated by $\widehat{A_tB_u}$, where

$$\widehat{A_tB_u} = \sum_1^q c_j\hat{A}_{t(j)} . \quad (9b)$$

The sum of squares of a product interaction contrast is computed from

$$SS_{A_tB_u} = \frac{n(\widehat{A_tB_u})^2}{(\sum_1^p c_i^2)\cdot(\sum_1^q c_j^2)}, \quad (10)$$

where c_i = the ith coefficient in the vector $\mathbf{A}_t$, and c_j = the jth coefficient in the vector $\mathbf{B}_u$. The degrees of freedom for a product interaction contrast are equal to

$$df_{A_tB_u} = (df_{A_t})\cdot(df_{B_u}), \quad (11)$$

where df_{A_t} = the degrees of freedom of the A_t contrast, and df_{B_u} = the degrees of freedom of the B_u contrast. Since the A_t and B_u treatment contrasts always have 1 degree of freedom each, an interaction contrast always has 1 degree of freedom.

For example, from Equation 9b, the estimate of the A_1B_1 contrast is $\widehat{A_1B_1} = (3)(.3) + (-1)(9.3) + (-1)(17.5) + (-1)(28.9) = -54.8$. From Equation 10, the sum of squares is $SS_{A_1B_1} = [(6)(-54.8)^2]/[(2)(12)] = 750.76$. The F ratio is $F_{A_1B_1} = (750.76/1)/19.15 = 39.20$. The Scheffé critical value for the 1 and 60 degrees of freedom F ratio is $S = 6F(6, 60) = 6(3.12) = 18.72$, $\alpha = .01$. Since the observed F of 39.20 is larger than 18.72, the A_1B_1 interaction contrast is judged significant. It can be concluded that the difference between the a_2 therapy group and the control group is larger under the drug conditions than under the placebo condition. In other words, the antianxiety drug increases the effectiveness of the a_2 therapy, when the effectiveness is measured by the difference between the control and therapy conditions.

Partial interactions can also be examined by starting with a B_u rather than an A_t treatment contrast. For example, earlier it was shown that the B_1 contrast [where $\mathbf{B}'_1 = (3\ -1\ -1\ -1)$] is significant. That is, on the average, the EDR is smaller under the drug conditions than under the placebo condition. It is of interest to determine whether this difference is the same for all three therapy groups. Simple Treatment B contrasts allow us to estimate the difference separately for each therapy group. A simple Treatment B contrast is denoted by $B_{u(i)}$ and defined as

$$B_{u(i)} = \sum_1^q c_j\mu_{ij}, \quad (12a)$$

where c_j = the jth coefficient in the vector $\mathbf{B}_u$. The $B_{u(i)}$ simple contrast is estimated by $\hat{B}_{u(i)}$, where

$$\hat{B}_{u(i)} = \sum_1^q c_jM_{ij} . \quad (12b)$$

For the B_1 contrast, the simple treatment contrast estimates are

$$\hat{B}_{1(1)} = 9.20,$$
$$\hat{B}_{1(2)} = 64.00,$$

and

$$\hat{B}_{1(3)} = 28.00 .$$

The sum of squares of an AB_u partial interaction can be computed from an equation analogous to Equation 6:

$$SS_{AB_u} = n\{\sum_1^p \hat{B}_{u(i)}^2 - [\sum_1^p \hat{B}_{u(i)}^2]/p\}/(\sum_1^q c_j^2) . \quad (13)$$

The degrees of freedom for an AB_u partial interaction is equal to

$$df_{AB_u} = (df_A)\cdot(df_{B_u}) . \quad (14)$$

Employing Equations 13 and 14, the sum of squares, degrees of freedom, and mean square of the AB_1 partial interaction are $SS_{AB_1} = 771.04$, $df_{AB_1} = 2$, and $MS_{AB_1} = 385.52$. The F ratio is therefore $F_{AB_1} = 385.52/19.15 = 20.13$. From Equation 8, the critical STP value is $G_2 = [6F(6, 60)]/2 = (6)(3.12)/2 = 9.36$, $\alpha = .01$. Since the observed F of 20.13 exceeds 9.36, the AB_1 partial interaction is judged significant. It is concluded that the difference between the EDR under the placebo condition and the average EDR under the three drug conditions is not the same for all therapy groups.

The precise nature of the AB_1 partial

Table 3
Summary of Analysis of Variance

Source	SS	df	MS	F
Treatment A	2,444.16	2	1,222.08	63.82*
A_1	2,352.00	1	2,352.00	122.88*
A_2	1,060.32	1	1,060.32	55.37*
A_3	253.92	1	253.92	13.26
Treatment B	2,370.96	3	790.32	41.27*
B_1	1,734.00	1	1,734.00	90.55*
Treatment A × Treatment B	1,376.40	6	229.40	11.98*
A_1B	1,332.12	3	444.04	23.19*
A_1B_1	750.76	1	750.76	39.20*
A_2B	175.98	3	58.66	3.06
AB_1	771.04	2	385.52	20.13*
Within cell	1,149.00	60	19.15	

Note. $A'_1 = (1\ \ -1\ \ 0)$; $A'_2 = (1\ \ 0\ \ -1)$; $A'_3 = (0\ \ 1\ \ -1)$; $B'_1 = (3\ \ -1\ \ -1\ \ -1)$. Total *SS*s = 7,340.52; total *df*s = 71.
* $p \leq .01$.

interaction can be determined by testing interaction contrasts. As in Equations 9a and 9b, an A_tB_u interaction contrast and its estimate, $\widehat{A_tB_u}$ can be defined as

$$A_tB_u = \sum_1^p c_iB_{u(i)}\,, \tag{15a}$$

and

$$\widehat{A_tB_u} = \sum_1^p c_i\hat{B}_{u(i)}\,, \tag{15b}$$

where c_i = the ith coefficient in the vector $\mathbf{A}_t$. The coefficient vector $\mathbf{A}_t$ describes the contrast among the $B_{u(i)}$ simple treatment contrasts.

An examination of the $B_{1(i)}$ simple contrast estimates suggests that the difference between the placebo and drug conditions is larger for therapy a_2 than for the control condition. The coefficient vector associated with this contrast is $A'_1 = (1\ \ -1\ \ 0)$. From Equation 15b, $\widehat{A_1B_1} = (1)(9.2) + (-1)(64.0) = -54.8$. Notice that this is the same result earlier obtained from Equation 9b. Equations 10 and 11 could be used to calculate the sum of squares and degrees of freedom for the interaction contrast. Of course, the results would again be the same. Table 3 presents an ANOVA table that summarizes all the treatment contrasts, partial interactions, and interaction contrasts tested.

Examples of partial interactions and interaction contrasts for a three-treatment design are provided by Boik (1975), who also describes alternative methods for controlling the Type I error rate when testing partial interactions and interaction contrasts.

Reference Note

1. Games, P. A. *Nesting, crossing, and the role of statistical models.* Paper presented at the meeting of the American Educational Research Association, New York, April 1977.

References

Boik, R. J. Interactions in the analysis of variance: A procedure for interpretation and a Monte Carlo comparison of univariate and multivariate methods for repeated measures designs (Doctoral dissertation, Baylor University, 1975). *Dissertation Abstracts International*, 1975, *36*, 2908B. (University Microfilms No. 75-27, 837)

Boik, R. J. A note on the rationale of Scheffé's method and the simultaneous test procedure. *Educational and Psychological Measurement*, 1979, *39*, 49–56.

Gabriel, K. R. A procedure for testing the homogeneity of all sets of means in analysis of variance. *Biometrics*, 1964, *20*, 459–477.

Gabriel, K. R. Simultaneous test procedure—Some theory of multiple comparisons. *Annals of Mathematical Statistics*, 1969, *40*, 224–250.

Games, P. A. Type IV errors revisited. *Psychological Bulletin*, 1973, *80*, 304–307.

Glass, G. V, & Hakstian, A. R. Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, 1969, *6*, 403–414.

Kirk, R. E. *Experimental design: Procedures for the behaviorial sciences.* Monterey, Calif.: Brooks/Cole, 1968.

Levin, J. R., & Marascuilo, L. A. Type IV errors and interactions. *Psychological Bulletin*, 1972, *78*, 368–374.

Levin, J. R., & Marascuilo, L. A. Type IV errors and Games. *Psychological Bulletin*, 1973, *80*, 308–309.

Marascuilo, L. A., & Levin, J. R. Appropriate post hoc comparisons for interactions and nested hypotheses in analysis of variance: The elimination of Type IV errors. *American Education Research Journal*, 1970, *7*, 397–421.

Marascuilo, L. A., & Levin, J. R. The simultaneous investigation of interaction and nested hypotheses in two-factor analysis of variance designs. *American Educational Research Journal*, 1976, *13*, 61–65.

Scheffé, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, *40*, 87–104.

Received May 8, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1090–1097

On Getting Good Subject Mileage: Reuse of Subjects in Experiments Involving Groups

John M. Light
Department of Sociology
Princeton University

Jerald Schutte
Department of Sociology
Columbia University

This article presents a combinatorial solution to the problem of reusing subjects, which often arises in experiments on groups. Blocking and Latin square approaches are examined and contrasted. A compromise solution is offered applying a variation of "v, k, λ" balancing techniques to incomplete block designs. Suggestions are offered for testing individual effects as well as for using ordinary one-way analysis of variance.

Experimental studies involving interacting groups are an omnipresent feature of social psychological literature. In some cases, researchers are able to simulate interaction using confederates, fictitious people, computers, and the like. But many substantive areas of social psychology require that actual groups be studied, for example, group polarization (Meyers & Lamm, 1976), interaction and attraction (Insko & Wilson, 1977), bystander intervention (Wegner & Schaefer, 1978), obedience to authority (Hamilton, 1978), and certainly many others. Sometimes the group can actually be the unit from which measurements of various dependent variables are taken, for example, studies of cooperation rates in Prisoners' Dilemma games (Dawes, McTavish, & Shaklee, 1977).

One aspect of group-level experiments poses an especially difficult problem. Suppose we are going to perform a simple one-factor completely randomized experiment on a set of subjects. Five or six people per treatment may be sufficient for the discovery of any treatment effects. But suppose that instead of people, the unit of analysis is groups, for example, groups of four. Then the researcher would need five or six groups per treatment or between 20 and 24 people. It should be clear that if the experiment consists of more than a few treatment levels, or the researcher complicates the design in any way (e.g., by adding another factor), many subjects may be required.

Even assuming that enough subjects are available to carry out such a design, the researcher still must contact, schedule, and pay these additional subjects. Therefore, with all elements considered, the typical group-level experiment is logistically more difficult to carry out than an experiment in which people themselves are the unit of analysis.

This being the case, any enterprising investigator would want to determine whether there is some legitimate way to reuse subjects in group-level experiments. In this article, we consider past approaches to this problem and then present a plan that we feel is a sensible alternative.[1]

Randomized Block Designs

Techniques for the reuse of subjects have typically been based on (or related to) ran-

This article is a collaborative effort and each author contributed equally.

Requests for reprints should be sent to John M. Light, Department of Sociology, Princeton University, Princeton, New Jersey 08540.

[1] Greenwald (1976) has considered the reuse of subjects in individual-level experiments in an article that is particularly interesting for its comments on the psychological aspects of reuse. Many of his points apply to our dsicussion of group-level experiments as well, and we refer to that article when appropriate.

domized block (RB) designs. In such designs, one ordinarily assembles a number of "blocks" of size t, where t is the number of experimental treatments (or treatment combinations), and assigns one member of each block at random to each treatment (Table 1). If the assumptions of the design are met, one can calculate treatment effects as well as subject (block) effects. The linear model for the design is

$$X_{ij} = \mu + \beta_j + \tau_i + \epsilon_{ij}, \quad (1)$$

where β_j is the effect of being in the jth treatment, and τ_i is the effect of being the ith block. Since the blocks are assembled so that they are relatively homogeneous with respect to the dependent variable under study, treatment effects can be more easily discovered (though not always, since one loses degrees of freedom in attempting to incorporate block effects). Of course, blocking can be used as one aspect of more complex designs (Kirk, 1968, p. 11).

With repeated measures designs, a block may simply correspond to one experimental subject on whom more than one of the treatment levels is applied; thus the subjects are reused. In the extreme case, the researcher can administer each treatment to each subject. There is an additional bonus in this case, since subjects act as their own controls (Kirk, 1968, p. 131).

Such a repeated measures procedure is probably the ideal way to employ the RB design. On the other hand, repeated measures can only be used in the RB design if administration of one treatment does not affect the results that would be expected by administration of a later treatment. With repeated measures on nonreactive units of analysis, such as rocks or leaves (as in natural science experiments), all treatment conditions are properly independent. But if human subjects are used, such independence is unlikely.[2]

Violation of the independence assumption can occur for two reasons. First, there may be some time-order effect, that is, subjects react differently to later trials than earlier ones. For example, suppose that a researcher is studying the extent to which people are persuasive under several different motivational conditions; subjects may improve their persuasion just by practicing on earlier trials (Greenwald, 1976, p. 316). Second, there may be some interaction effects; for example, a given treatment is reacted to differently depending on the treatments preceding it.[3]

In the event that a block repeated measures design is used for groups (i.e., each block is composed of one group that is repeatedly measured on the dependent variable), both time-order and Time-Order × Treatment effects should be more pronounced, since it is well-known that groups tend to negotiate definitions of situations (Asch, 1956; Festinger, 1950; Sherif, 1936).

Note that if treatments are randomized within blocks, the main effect of time order cancels out in treatment comparisons. It does, however, contribute to the error term and therefore reduces the power of the design. From this perspective, a repeated measure group-level RB design will tend to be conservative, though relatively inefficient (we reject fewer null hypotheses than we

Table 1
Randomized Block Design

Block	Treatment 1	2	3	4	5
1	0_1	0_2	0_3	0_4	0_5
2	0_1	0_2	0_3	0_4	0_5
3	0_1	0_2	0_3	0_4	0_5
4	0_1	0_2	0_3	0_4	0_5

Note. Each 0 represents one observation on a member of that block.

[2] Actually, the statistical requirement for an unbiased F test in a repeated measures design is that the Treatment × Treatment covariance matrix have equal off-diagonal elements (Winer, 1971, p. 277). This condition is satisfied if treatment covariances are only a result of measurements for each pair of treatments containing members of the same blocks (cf. Kirk, 1968, p. 140).

[3] All three types of context effects discussed by Greenwald (1976) are basically Time × Treatment interaction. His discussion of such effects is more detailed and contains more examples than ours. Much of that discussion applies to cases in which entire groups are reused as well.

Table 2
Latin Square Design

Subject	Time order 1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Note. Each letter represents a different treatment.

should). Time × Treatment interaction can similarly be removed by more elaborate randomization. But other interactions cannot and may either further obscure actual results or create spurious ones.

Latin Square Designs

One way to increase the power of an RB repeated measures design and to deal with the main effects of time order is to have each subject take the treatments in different time orders. In this way, each treatment is given first during one round, second during another, third during another, and so on. This should allow the researcher to extract these time-order effects from the error term, thereby increasing the power of the test.

In fact, this is just the approach taken in a Latin square (LS) design (Table 2). For a repeated measures LS design, we interpret the rows of the square as observational units (people, groups, etc.), the columns as the different times, and the table entries as treatments. The treatments are assigned so that each treatment appears only once in each row and column. In this way, each subject takes each treatment only once, and each treatment is taken at each possible different time only once. This is precisely the balancing condition previously described as desirable. The linear model for the LS design is

$$X_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{m(ijk)}, \quad (2)$$

where μ is the grand mean, α_i is the effect of row i (subject), β_j is the effect of column j (time), γ_k is the effect of treatment k, and ϵ is the experimental error.

At first, one might think that this design solves our problems. Certainly the major problem of controlling for time-order effects has been taken care of. At least a treatment effect is not suppressed by including the main effect of time order in the error term.

However, this design does not deal with interaction effects. It must be assumed that there are no interactions between time and treatment, time and subject, or treatment and subject (nor can there be any three-way interaction). If any exist, they will be confounded with treatment effects (in the case of Time × Subject interaction) or experimental error (in the case of Time × Treatment or Treatment × Subject).[4] Since the error term appears in the denominator of the F statistic for all main effects, occurrence of these last two types of interaction produces a conservative test (Cochran & Cox, 1957). On the other hand, if there is reason to believe that such interaction effects will be present, this conservative aspect of the LS design may be inefficient—a strong treatment effect is the only sort that would be detected.

Although Time × Subject interaction is confounded with treatment effects, if one is dealing with groups as the unit of analysis, it is probable that this type of interaction will not be a problem. In the case of Time × Group interaction, we are postulating that groups learn differently: Time order has different effects on different groups. This is an unlikely possibility. Moreover, it does not seem likely that groups will have peculiar responses to particular treatments. Thus Treatment × Group interaction should be an unlikely problem as well.

Both of these generalizations are based on the fact that groups tend to be more homogeneous than a comparable number of single individuals, since averaging of many individ-

[4] Evidently many people believe that all of these interactions will be confounded with main treatment effects, but this is not so. Interactions involving treatment (Time × Treatment or Subject × Treatment) will increase variance about the treatment mean, thus adding to error and obscuring real treatment effects. Only Time × Subject interaction will create between-means variance that artificially increases apparent treatment effects.

ual properties is possible. The larger the group, the more likely this is to be true. Although there are experiments in which the characteristics of particular individuals in each group may be important (e.g., problem solving, in which individual insight is the key to the solution), many if not most group experiments can rely on this homogenizing effect to create relatively similar-acting groups.

But Time × Treatment interaction (Greenwald, 1976) does pose a problem, much as it does with RB repeated measures designs. To say that such interaction exists is to say that a treatment has a different effect at a particular time. For example, it may be that in a bargaining game, large payoffs have more impact when administered in an early or late session; in the first case, there might be a primacy effect, and in the second, subjects would have had a chance to develop enough perspective to realize the magnitude of the payoff. As argued earlier, this could well be a problem when repeated measure is taken on individuals; it will be even worse when groups are the unit measured. The reason this is such a problem for group repeated measures designs, whether RB or LS, again rests on the fact that groups tend to be a source of reality conformation. This being the case, we would expect that group members' orientations toward a given treatment would be solidified in the group context, the more often that group context is reused. This would tend to exaggerate the Time × Treatment interaction. An individual alone would perhaps be less willing to make such judgments.

Besides being weakened by the presence of this interaction effect (an effect that we have argued will often be present), LS designs, whether applied to individuals or groups, suffer from another, even more basic, problem. Kirk (1968, p. 158) notes that it is impractical to use LS designs of dimension less than 5 × 5 with only one observation per cell (as in Table 2), since there are not enough degrees of freedom to give the design any power. This gives the researcher two choices: to replicate the square (thereby doubling the number of experimental units and to some extent defeating the purpose of the repeated measures procedure) or to use a fairly large square. But if the latter course can be followed (i.e., the experimenter is able to vary the number of treatment levels), fatigue effects can still ruin the experiment. Obviously, subjects' interest in the treatments will wane as they take more and more treatments. This may produce motivation to find a single mode of responding to all treatments, a strategy that would be exacerbated in the group context. As a result, if groups are given many treatments (as in a sizable LS design), later treatments will often produce homogeneous responses and are hence uninformative. In many cases, it is doubtful that the added degrees of freedom gained in the larger LS could make up for this increased homogeneity of response, and the treatment effects could again be lost.

A New Method for Reusing Subjects

We have argued that for logistic reasons of recruiting and scheduling as well as for statistical reasons of intersubject variability, the reuse of subjects is especially beneficial. But we have also noted that the typically used RB and LS repeated measures designs are likely to be unsatisfactory. The more similar each treatment situation is to previous and succeeding ones, the more the responses will tend to be interdependent. Moreover, we have argued that these dependencies should be even stronger with groups, since the group context offers members the opportunity to create some type of joint perspective as to the proper mode of orientation and behavior.

Therefore, we would like to propose a different statistical approach, which takes advantage of the fact that groups are the unit of analysis. Since reuse of the same group in repeated measures designs often creates the problem of interaction effects, we suggest a format that will systematically rotate people through different groups but that will do so in such a way that no two people will ever be in the same group more than once. In other words, we will arrange to guarantee that on each trial, a subject will receive the experimental treatment with

an entirely new and uniquely constituted group.

We thus make use of the fact that a large part of the context for each experimental treatment is a subject's group. By reusing subjects but not groups, we make certain that this important context is entirely novel on each trial. Under these conditions, it is likely that subjects will treat the trials as largely independent. Although statistically, individuals are still subject to sequence or carry-over effects, reconstituting the unit of analysis (i.e., the groups) each time assures that we minimize the problem of intertrial dependencies, particularly Time-Order × Treatment interaction.

The format for our approach stems from Yates (1936), who developed a class of designs that he saw as one solution to the restrictions inherent in randomized block designs (i.e., that one must have the same number of subjects per block as number of treatments). In such designs, each block is repeatedly measured but on fewer than all treatment levels. Yates called this procedure an "incomplete block" design to indicate that members of the block are given an incomplete subset of treatment levels of the independent variable.

This class of designs can be described as a combinatorial problem: How many ways can some $k < t$ treatment levels be assigned to blocks (individuals in a repeated measures design), and how many blocks will suffice. The binomial coefficient (t choose k) specifies the number of blocks (b) needed if all possible selections of k from t treatments are constructed, with the proviso that each treatment level is replicated r times and that any two treatments are paired together in a given block λ times.

Combinatorial problems with these constraints are often used in conjunction with creating balanced incomplete block (BIB) designs, such that the number of treatments remains constant across blocks. These designs, called (t, k, b, r, λ) configurations, are formally defined as a family of b subsets of a set S consisting of t elements, such that for some fixed k and λ, each subset has k elements, and each pair of elements of S occurs together exactly λ times. Given these assumptions, the following equations must hold (Fisher, 1940):[5]

$$N = rt = bk \tag{3}$$

$$r(k-1) = \lambda(t-1). \tag{4}$$

For example, consider the set of seven elements to be S = {1, 2, 3, 4, 5, 6, 7}. Now suppose we specify the following subsets of the power set of S: {1, 2, 4}, {2, 3, 5}, {3, 4, 6}, {4, 5, 7}, {5, 6, 1}, {6, 7, 2}, {7, 1, 3}. Here $b = 7$, $t = 7$, $k = 3$, and $\lambda = 1$; any two numbers appear in the same subset only once.

The matrix summarizing these sets is called an incidence matrix. In our example, the incidence matrix is

1101000
0110100
0011010
0001101
1000110
0100011
1010001

The number of rows is equal to the number of subsets b, and the number of columns is equal to the number of elements t. The number of entries in the columns is k. The number of entries in the rows is r.

In the language of experimental designs that use incomplete blocks, we have the following equivalences: t = treatment levels in the experiment[6] (the number of columns);

[5] This is easily proven. Since k and λ are fixed, and t is constant for any given experiment, r must be the same for each treatment. Hence, there are rt appearances of the treatments altogether. But there are b blocks, each of which receives k treatments. This must also equal the total number of appearances. Thus $bk = rt = N$. Further, since any one treatment level may appear exactly λ times with any other of the remaining $(t-1)$ treatment levels, it follows that the total number of times a treatment pair containing any one treatment will occur is $\lambda(t-1)$. But there are $(k-1)$ other treatment levels in any treatment's subset that appear in exactly r replications. This also equals the total number of times a treatment pair will occur. Hence, $r(k-1) = \lambda(t-1)$.

[6] A notational problem arises here. Originally, these problems were called "v, k, λ" designs; but more

b = the number of blocks (the number of rows); k = the number of treatment levels assigned to blocks (the number of 1s in each row); r = the number of replications of each treatment level (the number of 1s in each column); λ = the number of times any two treatment levels appear together in a block. For a design to be balanced, Equations 3 and 4 must hold, and t, b, k, r, and λ are integers. That this is true implies that k must be less than t, that is, each block must necessarily be assigned less than all of the treatments. It is this requirement that defines the block design as incomplete.

We now deal with a subset of the balanced incomplete block designs and demonstrate this subset's utility in group experiments (Table 3). The subset can be described with the following constraints. First, only those designs in which $\lambda = 1$ are dealt with. This consists of the subset of designs in which each treatment appears with any other treatment only once. Second, the designs are further restricted by assuming that these blocks are individuals, such that we essentially have an incomplete repeated measures procedure. Note that each column (Table 3) contains the same number of people and that each pair of people appears in the same column only once in the experiment. If we interpret the columns as being groups in addition to being treatments, then certainly for the example in Table 3, groups are being uniquely reconstituted each time, in accordance with our earlier desiderata. The following simple proposition shows that this will always be true when $\lambda = 1$ and (within the context of v, k, λ designs) will be true in only that case. It can be thought of as evidence for a type of "duality"; in this class of BIB designs, limiting the number of times that treatments appear together in the same block also restricts the number of times that people appear together.

recently, t has been used instead of v to denote the number of treatment levels. So we have used the term "v, k, λ" to refer to the class of problems, but we use t instead of v to denote the number of treatments, since most books on experimental design use such notation.

Table 3
Balanced Incomplete Block Design

	Treatment		
Block	1	2	3
1	0_1		0_2
2		0_1	0_2
3	0_1	0_2	

Note. Each 0 is one observation from the row block.

Proposition

In a v, k, λ configuration, $\lambda = 1$ if and only if no two people appear together in the same group more than once.

Proof. Let i and j be integers 1 thru t inclusive, let L and m be integers 1 thru b inclusive, and let $\mathbf{A}$ be the incidence matrix. Assume $\lambda = 1$, and suppose that two people appear together more than once. This implies that $\mathbf{A}_{Li} = \mathbf{A}_{mj} = 1$ for some pair of subjects L and m and some treatments i and j. But $\lambda = 1$ implies that $\mathbf{A}_{Li} = \mathbf{A}_{Lj} = 1$ for one and only one L, a contradiction. Now assume that no two people appear together more than once, and assume $\lambda = n > 1$. This means that n integers L_p exist, such that for $i = j$, $\mathbf{A}_{L_pi} = \mathbf{A}_{L_pj} = 1$ for all n values of L. But this means that two people, for example L_1 and L_2, see each other at least twice, in treatments i and j, a contradiction. This completes the proof.

It is interesting that by restricting our attention to v, k, λ designs in which $\lambda = 1$, we can ensure the added advantage that (to use our interpretation) no pair of subjects meets more than once in the course of the experiment. In this way, uniqueness of each experimental situation can be guaranteed, thereby discouraging subjects from treating later treatments like previous ones. We call this class of designs group balanced incomplete block (Group BIB) designs.

In addition to the formal consideration shown in the proposition, there are several practical considerations that we find limit the researcher's choice of group BIB designs. First, if we were to let λ be other than 1, then people could meet more than once during the course of the experiment. Although this might be acceptable in some types of

Table 4
Several Group Balanced Incomplete Block Designs

Design parameters				
t	r	b	k	λ
3	2	3	2	1
4	3	6	2	1
5	4	10	2	1
6	5	15	2	1

Note. Each row represents one design.

experiments, clearly our desire to make each experimental group maximally different causes us to prefer uniquely constituted groups. Hence, for this study, we restrict our attention to designs in which $\lambda = 1$. Second, observe that k must be at least 2, since otherwise subjects are not reused. But since $\lambda = 1$, if k is larger than 2, an inordinately large number of treatments are required to create a group BIB design (7 at a minimum and 13 or higher after that; see Table 11.3, Cochran & Cox, 1957, p. 469). Further restricting our attention, then, to cases in which $k = 2$, it follows from Equations 3 and 4 that $r = t - 1$; that is, the size of the group will be one less than the number of treatments.

Besides the practical benefit of keeping k set at 2, there is a possible substantive benefit: We are assured that any individual will receive the minimum possible number of treatments (while still being reused), thus minimizing the possibility of fatigue and/or learning effects. This is discussed more extensively later.

Since λ and k are restricted to be constants by these considerations, it follows again from Equations 1 and 2 that for practical purposes (i.e., for values of t at or below 7), the entire design is specified by one parameter t. If a researcher knows the number of treatment levels to be given, both the group size r and the number of subjects needed b can be determined. Table 4 enumerates some of these designs.

Advantages of the Design

Such subsets of incomplete block designs are important for several reasons. First, it can be observed that the number of subjects needed to produce one group observation in each treatment (i.e., b) will be considerably less than the minimum number of subjects needed for a RB or LS design, even when this incomplete design is reconstituted more than once to achieve multiple observations per treatment level.

Second, if the group measure is taken on the individual and aggregated, our design allows a row effect (i.e., subject effect) to be computed and used to adjust the treatment mean as in any incomplete block design. That is,

$$X_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}, \qquad (5)$$

where the estimate of β_i is used to adjust the estimate of τ_j. This means we can account for the effects of an individual being involved in a repeated measure and use this information to help assess the treatment effects.

More importantly, however, when the measure is at the group level, this procedure minimizes the problem inherent in assuming intertrial independence. Since subjects take more than one treatment, such independence will not hold in general. However, as previous arguments suggest, the group BIB design is explicitly arranged to minimize violation of this assumption under a usefully wide set of circumstances. Thus we think of the statistical analysis of a group BIB design as reducing to a one-way analysis of variance (ANOVA) in which each level of the treatment variable has one (group) observation. Each such design, conducted with new subjects, constitutes an additional observation per treatment level. Use of an ordinary one-way ANOVA on such data is justified because each group (i.e., unit of analysis) is independent and uniquely constituted, and each individual appears in the minimum possible repetitions (2).

As we suggested earlier, the fact that subjects have not seen each other before should create novel contexts that discourage use of information from previous trials. Further, no subject will ever have more than one such previous trial. Naturally, in experiments in which learning is an important variable (e.g.,

in studies of group problem solving in which problems are similar and in which time might be the dependent variable), we would not expect this kind of argument to hold. Generally, when the treatments are more the focus than the group context, there are likely to be carry-over effects. If the group context is the focal point for the subjects, however (with treatments more in the background, such as conditions for negotiations, etc.), we would expect group interaction to be less affected, and the approach presented here may be helpful.

These advantages provide intuitive and partial statistical grounds for assuming that such observations are independent. This eliminates much of the need to worry about time-order and various interaction effects, as in RB and LS designs.

Conclusion

We have discussed the disadvantages of doing group-type experiments by the usual methods—completely randomized factorial designs—as compared with attempting to reuse subjects in randomized block repeated measures or Latin square designs with time order as one of the nuisance variables. We believe our approach offers a good compromise between these techniques. By allowing subjects to be reused only in certain restricted ways, we have created conditions in which the usual problems of repeated measures designs should be minimized. At the same time, the researcher avoids the (often prohibitive) costs associated with recruiting and scheduling enough subjects to fill a completely randomized design.

The solution we have suggested is a modest one. On the other hand, it is important to see that its weaknesses differ from the weaknesses of other kinds of repeated measures designs; thus the group balanced incomplete block design should be thought of as another tool in the researcher's toolbag, one which will be useful only some of the time. We believe that when it is applicable, though, it may simplify experimental procedures significantly, enough to warrant being considered for use in any experiments on groups.

References

Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs,* 1956, *70*(9, Whole No. 416).

Cochran, W. G., & Cox, G. M. *Experimental designs.* New York: Wiley, 1957.

Dawes, R. M., McTavish, J., & Shaklee, H. Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology,* 1977, *35,* 1–11.

Festinger, L. Informal social communication. *Psychological Review,* 1950, *57,* 271–282.

Fisher, R. A. An examination of the different possible solutions of a problem in incomplete blocks. *Annals of Eugenics,* 1940, *1,* Pt. 2, 52–75.

Greenwald, A. G. Within-subjects designs: To use or not to use? *Psychological Bulletin,* 1976, *83,* 314–320.

Hamilton, V. L. Obedience and responsibility: A jury simulation. *Journal of Personality and Social Psychology,* 1978, *36,* 126–146.

Insko, C. A., & Wilson, M. Interpersonal attraction as a function of social interaction. *Journal of Personality and Social Psychology,* 1977, *35,* 903–911.

Kirk, R. E. *Experimental design: Procedures for the behavioral sciences.* Belmont, Calif.: Brooks/Cole, 1968.

Meyers, D. G., & Lamm, H. The group polarization phenomenon. *Psychological Bulletin,* 1976, *83,* 602–627.

Sherif, M. *The psychology of social norms.* New York: Harper, 1936.

Wegner, D. M., & Schaefer, D. The concentration of responsibility: An objective self-awareness analysis of group size effects in helping situations. *Journal of Personality and Social Psychology,* 1978, *36,* 147–155.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Yates, F. Incomplete Latin squares. *Journal of Agricultural Science,* 1936, *26,* 301–315.

Received May 12, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1098–1106

Comparison of Sequences

Lawrence J. Hubert
School of Education, University of California, Santa Barbara

The problem of comparing K numerical sequences, each defined over the same n objects, is approached through the use of a K-variate scoring function. Besides encompassing some old work of Spearman, the sequence-comparison paradigm can be used to discuss a number of separate data-analysis procedures of particular interest to the behavioral sciences, for example, nominal scale response agreement among multiple raters, Friedman's test, Kendall's coefficient of concordance, Page's test, and so on. The major intent of this article is pedagogical, and it emphasizes a general conceptual framework that conveniently organizes a number of well-known statistical strategies.

It should be apparent to every beginning student of statistics that many, if not most, data-analysis strategies are based on sums of squared differences. As is well-known, sums of this type appear in the definition of a variance, in least-squares regression, in chi-square goodness-of-fit tests, and so on. Upon reflection, however, a psychologist with even a superficial knowledge of test theory might consider this reliance on squared differences misplaced. As Spearman (1906) noted some 70 years ago, extreme scores are expected to contain relatively large amounts of measurement error; consequently, statistical procedures that accentuate these extremes by an additional process of squaring could easily compound the difficulties inherent in making inferences from fallible data. This last observation, along with some outdated problems of computational tractability, originally lead Spearman to consider an alternative measure of correlation, now called Spearman's footrule, based on a sum of absolute differences between two numerical sequences. In fact, Spearman proposed the footrule as a direct competitor to his more well-known rank correlation coefficient based on an analogous sum of squared differences.

Partial support of this research was supplied by National Science Foundation Grant SOC-77-28227.

Requests for reprints should be sent to Lawrence J. Hubert, Graduate School of Education, University of California, Santa Barbara, California 93106.

Although Spearman's contributions are now important primarily for historical reasons, this early work still represents a significant effort to develop alternative norms for the comparison of two numerical sequences. In particular, the present discussion is concerned with the more general problem of measuring the correspondence among K numerical sequences using an arbitrary scoring function, but I rely on the two simple norms presented by Spearman as an introductory example. Besides suggesting several interesting theoretical generalizations of Spearman's ideas, these extensions to multiple sequences lead to several new insights into the superficially different problem of measuring nominal scale response agreement among multiple raters (cf. Cohen, 1968; Fleiss, Cohen, & Everitt, 1969; Hubert, 1977). In addition, the same general structure is relevant to the problem of analyzing m rankings, for example, Friedman's two-way analysis of variance by ranks (Friedman, 1937), Kendall's coefficient of concordance (Kendall, 1970), and Page's L test (Page, 1963). It should be noted at the outset that this article's major contribution is pedagogical, and I do not intend to advocate any new data-analysis procedures. Inasmuch as all of the techniques mentioned here have been discussed in detail in the literature, there is no need to review the specifics of these individual data-analysis strategies. Apparently, the ubiquitous nature of the sequence-comparison problem has not been recognized in the methodological litera-

ture of psychology, and for this reason alone, an explicit discussion of the general paradigm may be of value to develop in greater depth.

Background

To introduce some terminology, suppose that $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ denote two numerical sequences in which the corresponding elements x_i and y_i are matched in some manner. Clearly, this framework is the natural context for developing a measure of association based on the n bivariate pairs $(x_1, y_1), \ldots, (x_n, y_n)$. When rank correlations are desired, an initial transformation of the xs and the ys to ranks may also be imposed, but for our purposes this reduction is unnecessary. At least for now, the original sequences of the xs and the ys can be retained.

Instead of proceeding directly to calculate a traditional Pearson product-moment correlation coefficient, suppose that a more general measure of proximity between the two sequences is of initial interest. In particular, let $f(\cdot, \cdot)$ be a bivariate function and define a summary measure of proximity between the sequences $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ by the index Γ:

$$\Gamma = \sum_{i=1}^{n} f(x_i, y_i).$$

Although the function $f(\cdot, \cdot)$ is arbitrary, two special cases are of obvious importance. If $f(x, y) = |x - y|$, then Γ forms the basis for Spearman's footrule; if $f(x, y) = (x - y)^2$, the basis for Spearman's more common rank correlation statistic is obtained. As a convention, Γ_α will refer to an index defined by $f(x, y) = |x - y|^\alpha$; thus, the footrule corresponds to the use of Γ_1 and Spearman's more common statistic corresponds to the use of Γ_2.

Although various normalizations of Γ may be desirable to transform a raw index into a suitably restricted measure of association, examination of these normalizations can be delayed until later and then discussed in greater generality. As it stands, the raw index Γ is sufficient for purposes of hypothesis testing in the permutation context; for example, the tables given in Kendall (1970) for the better known Spearman index of rank correlation are given in terms of Γ_2. To carry out such a test, an index of the form specified by Γ is evaluated under the notion of *independence*, in which all permutations of the ys against the fixed sequence of xs are considered equally likely (or equivalently, fixing the xs and permuting the ys). Thus, the exact null distribution of Γ can be obtained by evaluating the index over all $n!$ permutations of the ys and tabulating the resulting frequency distribution. If the index Γ is sufficiently extreme when compared to this distribution, the hypothesis of independence can be rejected.

Unless the xs and the ys are untied within their respective sets and a transformation to ranks (or to some other canonical form) is imposed, the permutation distribution has to be recalculated for each different application. Since this is a very extensive computational burden, large-sample (normal) approximations are desirable. As it turns out, these approximations are fairly easy to obtain through general formulas for the mean and variance of Γ, which can then be specialized for particular defining functions $f(\cdot, \cdot)$.

Γ as a Bilinear Permutation Statistic

The index Γ has the form of a bilinear permutation statistic (see Puri & Sen, 1971), and thus, assuming all permutations of the ys are equally likely,

$$E(\Gamma) = (1/n) \sum_{j=1}^{n} \sum_{i=1}^{n} f(x_i, y_j);$$

$$\operatorname{var}(\Gamma) = \{1/[n(n-1)]\}[(1/n)A_1 - (A_2 + A_3) + nA_4],$$

where

$$A_1 = [\sum_{j=1}^{n} \sum_{i=1}^{n} f(x_i, y_j)]^2;$$

$$A_2 = \sum_{j=1}^{n} [\sum_{i=1}^{n} f(x_i, y_j)]^2;$$

$$A_3 = \sum_{i=1}^{n} [\sum_{j=1}^{n} f(x_i, y_j)]^2;$$

$$A_4 = \sum_{j=1}^{n} \sum_{i=1}^{n} f(x_i, y_j)^2.$$

Specializing these formulae for the case in which the sequences $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$

represent the untied ranks of 1, 2, ..., n in some order, we have

Γ_1:

$$E(\Gamma_1) = (n^2 - 1)/3;$$
$$\mathrm{var}\,(\Gamma_1) = (n + 1)(2n^2 + 7)/45,$$

Γ_2:

$$E(\Gamma_2) = n(n^2 - 1)/6;$$
$$\mathrm{var}\,(\Gamma_2) = [1/(n - 1)][(n^3 - n)/6]^2.$$

The mean and variance of Γ_1 for untied ranks are available in Spearman (1906) and those for Γ_2 in Kendall (1970). Finally, under very mild conditions on the values assigned by the function $f(\cdot, \cdot)$, Γ can be shown asymptotically normal as $n \to \infty$. For a discussion of several possible sufficient conditions, the reader is referred to Puri and Sen (1971, p. 72) and Hoeffding (1951).

Correlations Among Indices

In addition to information regarding the single index Γ that can be obtained as shown above, it is relatively straightforward to calculate the correlation between two such indices, say Γ and Γ', based on two different bivariate functions $f(\cdot, \cdot)$ and $f'(\cdot, \cdot)$. First, the sum of $f(\cdot, \cdot)$ and $f'(\cdot, \cdot)$ is considered as a new bivariate function, and the variance of this index is obtained. The covariance for Γ and Γ' is then isolated by subtracting off the variances for Γ and Γ' and dividing by 2. Carrying out this process leads to the following general expression for the covariance between Γ and Γ' over the same $n!$ permutations:

$$\mathrm{cov}\,(\Gamma, \Gamma') = \{1/[n(n - 1)]\}[(1/n)B_1 - (B_2 + B_3) + nB_4],$$

where

$$B_1 = [\sum_{j=1}^{n}\sum_{i=1}^{n} f(x_i, y_j)][\sum_{j=1}^{n}\sum_{i=1}^{n} f'(x_i, y_i)];$$

$$B_2 = \sum_{j}[\sum_{i} f(x_i, y_j)\sum_{i} f'(x_i, y_j)];$$

$$B_3 = \sum_{i}[\sum_{j} f(x_i, y_j)\sum_{j} f'(x_i, y_j)];$$

$$B_4 = \sum_{j=1}^{n}\sum_{i=1}^{n} f(x_i, y_j)f'(x_i, y_j).$$

Finally, normalizing by the square root of the variances for Γ and Γ', the Pearson correlation between Γ and Γ' can be derived. As an example for the special case of $\Gamma \equiv \Gamma_1$ and $\Gamma' \equiv \Gamma_2$ and where untied ranks are used for the xs and ys, this procedure leads to the following simplification:

$$\rho(\Gamma_1, \Gamma_2) = \frac{3}{\sqrt{10}}\left(\frac{n^2 + 1}{\sqrt{(n^2 - 1)(n^2 + \frac{7}{2})}}\right).$$

Surprisingly, this correlation is bounded away from 1 for large n and approaches $3/\sqrt{10}$ as $n \to \infty$. This situation contrasts with an asymptotic correlation of 1.00 between Spearman's rank correlation index based on Γ_2 and Kendall's tau statistic (Kendall, 1970).

Although the details of index correlation will not be developed in any further detail, it is significant to note that such correlations are easily obtained once the variance of a general statistic, such as Γ, is derived (or the later extensions of Γ to multiple sequences). Thus, when a researcher is faced with alternative choices for an index, it may be of practical interest to know how highly the various choices intercorrelate under independence and for various values of n or, more specifically, to know if the intercorrelations are less than unity even when n is assumed infinite.

Other Applications of Γ

Besides the Spearman indices, the measure Γ also includes several other statistics of interest to psychology that have been developed in comparative isolation. For instance, as discussed by Hubert (1978), the index Γ can be used to develop Cohen's (1968) index of nominal scale response agreement between two raters. Here, the numerical sequence x_1, x_2, ..., x_n represents the labels for the R categories used by Rater 1 to classify n objects, and y_1, y_2, ..., y_n represents the labels used by Rater 2 to classify the same n objects into C categories. Typically, $R = C$ and the number of objects placed in the same category by the two raters is of interest. More generally, we can define

$$f(x_i, y_j) = w_{uv},$$

if x_i is placed in the category labeled u by Rater 1, $1 \leq u \leq R$, and in the category labeled v by Rater 2, $1 \leq v \leq C$. The index Γ

is then a raw index of weighted nominal scale response agreement that can be subjected to various normalizations to provide an appropriately restricted final index. As with the Spearman statistics, however, the raw index Γ can be considered by itself for the purposes of hypothesis testing. It should be apparent that an appropriate choice of weights (e.g., $w_{uv} = |u - v|^{\alpha}$) leads us directly back to the index Γ_α. Also, for Cohen's problem of nominal scale response agreement, the effect of different choices for the weights can be partially assessed by the general covariance formula between Γ and Γ' given earlier.

Multiple Sequences

The obvious extension of the index Γ to three sequences $x_1, \ldots, x_n$; $y_1, \ldots, y_n$; $z_1, \ldots, z_n$ relies on a trivariate function $g(x_i, y_j, z_k)$. As a notation, suppose the index associated with this function is given by Λ_3:

$$\Lambda_3 = \sum_i g(x_i, y_i, z_i), \tag{1}$$

and assume that all $n!$ reorderings of the ys and all $n!$ reorderings of the zs are equally likely under a hypothesis of independence. By carrying out the usual moment calculations (cf. Hubert, 1979),

$$E(\Lambda_3) = \frac{1}{n^2}\Big[\sum_{i,j,k} g(x_i, y_j, z_k)\Big];$$

$$\begin{aligned} V(\Lambda_3) = {} & \frac{2n-1}{n^4(n-1)^2}\Big[\sum_{i,j,k} g(x_i, y_j, z_k)\Big]^2 \\ & + \frac{n-2}{n(n-1)^2}\sum_{i,j,k} g(x_i, y_j, z_k)^2 \\ & - \frac{1}{n^2(n-1)^2}\Big\{\sum_i\Big[\sum_{j,k} g(x_i, y_j, z_k)\Big]^2 \\ & + \sum_j\Big[\sum_{i,k} g(x_i, y_j, z_k)\Big]^2 \\ & + \sum_k\Big[\sum_{i,j} g(x_i, y_j, z_k)\Big]^2\Big\} \\ & + \frac{1}{n^2(n-1)^2}\Big\{\sum_{i,j}\Big[\sum_k g(x_i, y_j, z_k)\Big]^2 \\ & + \sum_{i,k}\Big[\sum_j g(x_i, y_j, z_k)\Big]^2 \\ & + \sum_{j,k}\Big[\sum_i g(x_i, y_j, z_k)\Big]^2\Big\} \\ \approx {} & \frac{2}{n^5}\Big[\sum_{i,j,k} g(x_i, y_j, z_k)\Big]^2 \\ & + \frac{1}{n^2}\sum_{i,j,k} g(x_i, y_j, z_k)^2 \\ & - \frac{1}{n^4}\Big\{\sum_i\Big[\sum_{j,k} g(x_i, y_j, z_k)\Big]^2 \\ & + \sum_j\Big[\sum_{i,k} g(x_i, y_j, z_k)\Big]^2 \\ & + \sum_k\Big[\sum_{i,j} g(x_i, y_j, z_k)\Big]^2\Big\} \\ & + \text{lower order terms.} \end{aligned} \tag{2}$$

The general form for these moment expressions for K sequences follow in a similar manner. If $x_1^{(r)}, \ldots, x_n^{(r)}$, for $1 \le r \le K$, denote the K sequences, and if the index Λ_K is defined using a K-variate function $h[x_{i_1}^{(1)}, x_{i_2}^{(2)}, \ldots, x_{i_K}^{(K)}]$, then

$$\Lambda_K = \sum_i h[x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(K)}];$$

$$E(\Lambda_K) = \frac{1}{n^{K-1}}\sum_{i_1,\cdots,i_K} h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}];$$

$$\begin{aligned} \operatorname{var}(\Lambda_K) = {} & -\left\{\frac{1}{n^{K-1}}\sum_{i_1,\cdots,i_K} h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}]\right\}^2 \\ & + \frac{1}{n^{K-1}}\sum_{i_1,\cdots,i_K} h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}]^2 \\ & + \left[\frac{1}{n(n-1)}\right]^{K-1} \\ & \times\Big(\Big\{\sum_{i_1,\cdots,i_K} h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(k)}]\Big\}^2 \\ & - \sum_{i_1}\Big\{\sum_{i_2,\cdots,i_K} h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}]\Big\}^2 \\ & - \cdots - \sum_{i_K}\Big\{\sum_{i_1,\cdots,i_{K-1}} h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}]\Big\}^2 \\ & + \sum_{i_1,i_2}\Big\{\sum_{i_3,\cdots,i_K} h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}]\Big\}^2 \\ & + \cdots + \sum_{i_{K-1},i_K}\Big\{\sum_{i_1,\cdots,i_{K-2}} \\ & h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}]\Big\}^2 \\ & - \cdots - \cdots (-1)^K \\ & \times \sum_{i_1,\cdots,i_K}\{h[x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}]\}^2\Big) \end{aligned}$$

$$\approx\left(\frac{K-1}{n^{2K-1}}\right)\{\sum_{i_1,\cdots,i_K} h[x_{i_1}{}^{(1)},\ldots,x_{i_K}{}^{(K)}]\}^2 + \frac{1}{n^{K-1}}\{\sum_{i_1,\cdots,i_K} h[x_{i_1}{}^{(1)},\ldots,x_{i_K}{}^{(K)}]\}^2 - \frac{1}{n^{2(K-1)}}(\sum_{i_1}\{\sum_{i_2,\cdots,i_K} h[x_{i_1}{}^{(1)},\ldots,x_{i_K}{}^{(K)}]\}^2 + \cdots + \sum_{i_K}\{\sum_{i_1,\cdots,i_{K-1}} h[x_{i_1}{}^{(1)},\ldots,x_{i_K}{}^{(K)}]\}^2) + \text{lower order terms.}$$

Contingency Table Applications

As one particular application of the index Λ_k, suppose a threefold contingency table is given having R rows, S columns, and T layers, and let n_{rst} denote the number of observations in row r, column s, and layer t. If w_{rst} defines a fixed weight attached to the corresponding cell, then the raw index Λ_3 can be written as

$$\Lambda_3 = \sum_{r,s,t} w_{rst} n_{rst},$$

where the function $g(x_i, y_j, z_k)$ in Equation 1 is defined as w_{rst} if x_i belongs to row r, y_j belongs to column s, and z_k belongs to layer t. The moment formulas reduce in a similar manner; for instance,

$$E(\Lambda_3) = \frac{1}{n^2}\sum_{r,s,t} w_{rst} n_{r\cdot\cdot} n_{\cdot s\cdot} n_{\cdot\cdot t};$$

and using the approximate variance that ignores lower order terms,

$$\text{var}(\Lambda_3) \approx \frac{2}{n^5}(\sum_{r,s,t} w_{rst} n_{r\cdot\cdot} n_{\cdot s\cdot} n_{\cdot\cdot t})^2 + \frac{1}{n^2}(\sum_{r,s,t} w^2_{rst} n_{r\cdot\cdot} n_{\cdot s\cdot} n_{\cdot\cdot t}) - \frac{1}{n^4}[\sum_r n_{r\cdot\cdot}(\sum_{s,t} w_{rst} n_{\cdot s\cdot} n_{\cdot\cdot t})^2 + \sum_s n_{\cdot s\cdot}(\sum_{r,t} w_{rst} n_{r\cdot\cdot} n_{\cdot\cdot t})^2 + \sum_t n_{\cdot\cdot t}(\sum_{r,s} w_{rst} n_{r\cdot\cdot} n_{\cdot s\cdot})^2].$$

If we let $p_{r\cdot\cdot} = n_{r\cdot\cdot}/n$, $p_{\cdot s\cdot} = n_{\cdot s\cdot}/n$, and $p_{\cdot\cdot t} = n_{\cdot\cdot t}/n$, then

$$E(\Lambda_3) = n\sum_{r,s,t} w_{rst} p_{r\cdot\cdot} p_{\cdot s\cdot} p_{\cdot\cdot t};$$

and

$$\text{var}(\Lambda_3) \approx n[2(\sum_{r,s,t} w_{rst} p_{r\cdot\cdot} p_{\cdot s\cdot} p_{\cdot\cdot t})^2 + \sum_{r,s,t} w^2_{rst} p_{r\cdot\cdot} p_{\cdot s\cdot} p_{\cdot\cdot t} - \sum_r p_{r\cdot\cdot}(\sum_{s,t} w_{rst} p_{\cdot s\cdot} p_{\cdot\cdot t})^2 - \sum_s p_{\cdot s\cdot}(\sum_{r,t} w_{rst} p_{r\cdot\cdot} p_{\cdot\cdot t})^2 - \sum_t p_{\cdot\cdot t}(\sum_{r,s} w_{rst} p_{r\cdot\cdot} p_{\cdot s\cdot})^2].$$

When $R = S = T$, and three raters are assumed to define the dimensions of the three-way contingency table, a number of raw indices of rater agreement can be obtained by varying the definition of w_{rst}. As indicated by David and Barton (1962) and developed in the rater context by Hubert (1977), many different concepts of agreement are possible. I shall list several of these definitions later but give the specializations of the moments for only the first three since these appear to be the most natural.[1] The reader is referred to David and Barton (1962) for a discussion of asymptotic normality for these various alternatives when the weights are restricted to be dichotomous (i.e., $0 - 1$) or when they arise as sums of dichotomous weights.

1. DeMoivre: An agreement occurs if and only if all raters place an object in the same category. Thus, letting $w_{rst} = 1$ if $r = s = t$ and 0 otherwise, we obtain

$$E(\Lambda_3) = n\sum_r p_{r\cdot\cdot} p_{\cdot r\cdot} p_{\cdot\cdot r};$$

$$\text{var}(\Lambda_3) \approx n[2(\sum_r p_{r\cdot\cdot} p_{\cdot r\cdot} p_{\cdot\cdot r})^2 + \sum_r p_{r\cdot\cdot} p_{\cdot r\cdot} p_{\cdot\cdot r}(1 - p_{\cdot r\cdot} p_{\cdot\cdot r} - p_{r\cdot\cdot} p_{\cdot\cdot r} - p_{r\cdot\cdot} p_{\cdot r\cdot})].$$

Alternatively, if we let a_{rs} denote a weighting between Raters 1 and 2, b_{rt} a weighting

[1] It should be noted that these expressions correspond to those given in Hubert (1977) except for a typographical error in the first definition due to DeMoivre. Here, the sums over k should have been products over k; also, the last such product should be over $k' \neq k$ (this last change also points out a misprint in David and Barton, 1962). The correct formulas are given in the text of the present article.

between Raters 2 and 3, and c_{st} a weighting between Raters 1 and 3, w_{rst} = minimum (a_{rs}, b_{rt}, c_{st}). In the specific cases we are considering here, $a_{rs} = 1$ if $r = s$ and 0 otherwise, $b_{rt} = 1$ if $r = t$ and 0 otherwise, and $c_{st} = 1$ if $s = t$ and 0 otherwise. This notation will be used below as well.

2. Target: If the first rater is considered a target, then an agreement occurs if and only if another rater places an object in the same category as the first. More explicitly, let $w_{rst} = a_{rs} + b_{rt}$, then

$$E(\Lambda_3) = n \sum_{r,s,t} (a_{rs} + b_{rt}) p_{r\cdot\cdot} p_{\cdot s\cdot} p_{\cdot\cdot t}$$

$$= n[\sum_r p_{r\cdot\cdot}(p_{\cdot r\cdot} + p_{\cdot\cdot r})];$$

$$\operatorname{var}(\Lambda_3) \approx n\{2[\sum_{r,s,t} (a_{rs} + b_{rt}) p_{r\cdot\cdot} p_{\cdot s\cdot} p_{\cdot\cdot t}]^2$$

$$+ \sum_{r,s,t} (a_{rs} + b_{rt})^2 p_{r\cdot\cdot} p_{\cdot s\cdot} p_{\cdot\cdot t}$$

$$- \sum_r p_{r\cdot\cdot}[\sum_{s,t} (a_{rs} + b_{rt}) p_{\cdot s\cdot} p_{\cdot\cdot t}]^2$$

$$- \sum_s p_{\cdot s\cdot}[\sum_{s,t} (a_{rs} + b_{rt}) p_{r\cdot\cdot} p_{\cdot\cdot t}]^2$$

$$- \sum_t p_{\cdot\cdot t}[\sum_{r,s} (a_{rs} + b_{rt}) p_{r\cdot\cdot} p_{\cdot s\cdot}]^2\}$$

$$= n[\sum_r p_{r\cdot\cdot} p_{\cdot r\cdot}(1 - p_{\cdot r\cdot} - p_{r\cdot\cdot})$$

$$+ \sum_r p_{r\cdot\cdot} p_{\cdot\cdot r}(1 - p_{\cdot\cdot r} - p_{r\cdot\cdot})$$

$$+ (\sum_r p_{r\cdot\cdot} p_{\cdot r\cdot})^2 + (\sum_r p_{r\cdot\cdot} p_{\cdot\cdot r})^2].$$

3. Pairwise: An agreement occurs if and only if two raters categorize an object consistently. Thus, if we let $w_{rst} = a_{rs} + b_{rt} + c_{st}$, then

$$E(\Lambda_3) = n(\sum_r p_{r\cdot\cdot} p_{\cdot r\cdot} + \sum_r p_{r\cdot\cdot} p_{\cdot\cdot r}$$

$$+ \sum_r p_{\cdot r\cdot} p_{\cdot\cdot r});$$

$$\operatorname{var}(\Lambda_3) \approx n[(\sum_r p_{r\cdot\cdot} p_{\cdot r\cdot})^2 + (\sum_r p_{r\cdot\cdot} p_{\cdot\cdot r})^2$$

$$+ \sum_r p_{\cdot r\cdot} p_{\cdot\cdot r})^2$$

$$+ \sum_r p_{r\cdot\cdot} p_{\cdot r\cdot}(1 - p_{\cdot r\cdot} - p_{r\cdot\cdot})$$

$$+ \sum_r p_{r\cdot\cdot} p_{\cdot\cdot r}(1 - p_{r\cdot\cdot} - p_{\cdot\cdot r})$$

$$+ \sum_r p_{\cdot r\cdot} p_{\cdot\cdot r}(1 - p_{\cdot r\cdot} - p_{\cdot\cdot r})].$$

Although these three interpretations given for w_{rst} are probably the most obvious, several other possibilities exist:

4. If $w_{rst} = 1$ when at least one nontarget rater matches the target and 0 otherwise, then w_{rst} = maximum (a_{rs}, b_{rt}).

5. If $w_{rst} = 1$ when at least one pair of raters match and 0 otherwise, then w_{rst} = maximum (a_{rs}, b_{rt}, c_{st}).

6. If w_{rst} denotes the number of matches for pairs of raters with consecutive labels, then $w_{rst} = a_{rs} + c_{st}$.

7. If $w_{rst} = 1$ when a majority of raters match and 0 otherwise, then $w_{rst} = 1$ if $a_{rs} + b_{rt} + c_{st} \geq 2$ and 0 otherwise.

Although I shall not pursue these latter definitions or extensions to more than three raters (or in general, a G-fold contingency table for G greater than three), the approach follows exactly that given above and can be carried out by the reader.

Measuring Concordance in K *Rankings*

One of the standard nonparametric data-analysis problems is discussed under the title of "Friedman's Test" or "Kendall's Coefficient of Concordance." Here, K judges assign numerical values to n objects (e.g., ranks) and our interest is in (a) testing whether the n objects can be considered equally preferable and (b) measuring the degree of concordance among the K judges. In our context, the traditional approach to both of these problems relies on a particular K-variate function $h[x_{i_1}{}^{(1)}, \ldots, x_{i_K}{}^{(K)}]$ of the form

$$h[x_{i_1}{}^{(1)}, \ldots, x_{i_K}{}^{(K)}] = \sum_{k<k'} h[x_{i_k}{}^{(k)}, x_{i_{k'}}{}^{(k')}],$$

where

$$h[x_{i_k}{}^{(k)}, x_{i_{k'}}{}^{(k')}] = [x_{i_k}{}^{(k)} - x_{i_{k'}}{}^{(k')}]^2.$$

Thus, the raw index Λ_k can be written as

$$\Lambda_k = \sum_i (\sum_{k<k'} (x_i{}^{(k)} - x_i{}^{(k')})^2).$$

When viewed in a K by n analysis-of-variance format, Λ_k is merely K times the sum of squares *within* the n columns. Furthermore, if ranks are used rather than the original observations, implying that all row sums are equal, Λ_K is simply K times the sum of squares for interaction. Thus, since Λ_K/K and the sum of squares between columns must sum to the constant total sum of squares, a test statistic and a final normalized index could be defined equivalently using either of the former two quantities. Friedman's statistic and Kendall's coefficient of concordance W are defined in terms of the sum of squares between columns, but, as we will see later, W could be obtained just as well using Λ_K/K.

Viewed another way based on the Spearman norm Γ_2, denoted here by $\Gamma_{kk'}$ for the kth and k'th sequences,

$$\Lambda_K = \sum_{k<k'} \Gamma_{kk'},$$

and using the general formulas given previously,

$$E(\Lambda_K) = \sum_{k<k'} E(\Gamma_{kk'});$$

$$\text{var}(\Lambda_K) = \sum_{k<k'} V(\Gamma_{kk'}).$$

This result is also reflected in the fact that the Γ indices are independent in pairs (cf. David & Barton, 1962, p. 218), and suggests that very simple mean and variance formulas result when the function $h(x_{i_1}{}^{(1)}, \ldots, x_{i_K}{}^{(K)})$ is *additive* with respect to the appropriate bivariate functions. Thus, similar results would hold for $\Gamma_{kk'}$ defined using Γ_1 or any other bivariate function. Moreover, it is relatively simple to propose alternative *nonadditive* functions that measure variability within a column of the K by n table (e.g., the range) that would lead to alternative raw indices of concordance.

It should be apparent at this point that the problem of nominal scale response agreement among K raters can be rephrased in a generalized Friedman context involving K sequences of n observations. Instead of using ranks, however, category labels are attached to each of the n objects by each rater.

Testing for an a Priori Order in K *Rankings*

Instead of a general hypothesis test of the Friedman type, we can also define a test procedure sensitive to a particular a priori ordering of the n objects. Such an extension can be viewed as an analogue of the target-rater agreement problem mentioned earlier. If an a priori set of weights is given by the first sequence $x_1{}^{(1)}, \ldots, x_n{}^{(1)}$, then an appropriate test statistic can be given in the form used by Page (1963) and by Pirie and Hollander (1972) based on the function

$$h[x_{i_1}{}^{(1)}, \ldots, x_{i_K}{}^{(K)}] = x_{i_1}{}^{(1)} \sum_{k=2}^{K} x_{i_k}{}^{(k)}.$$

Thus, the index

$$\Lambda_K = \sum_{i=1}^{n} x_i{}^{(1)} \sum_{k=2}^{K} x_i{}^{(k)},$$

can be interpreted as a weighted sum of column totals, where each total is defined over the last $K-1$ sequences. Typically, the weights are integers from 1 to n, and the values $x_i{}^{(k)}$ for $k \geq 2$ are ranks (Page, 1963) or normal scores (Pirie & Hollander, 1972). The mean and variance formulas follow directly from the previous expressions, or alternatively, since Λ_K can be defined by the sum of the $K-1$ independent terms

$$\sum_{i=1}^{n} x_i{}^{(1)} x_i{}^{(k)}, \quad \ldots, \quad \sum_{i=1}^{n} x_i{}^{(1)} x_i{}^{(K)},$$

the mean and variance can be obtained by the usual moment formulas for linear combinations of random variables. Obviously, the formulas for the mean and variance of each individual term can be found by the relatively simple expressions used in the bivariate function examples based on two sequences.

Significance Testing

Although some proofs of asymptotic normality exists for various special cases of the index Λ_K (cf. David & Barton, 1962; Hoeffding, 1951), as well as proofs for the convergence to a chi-square random variable for related statistics such as Friedman's (see Lehmann, 1975), all these approximations are of varying adequacy depending on the size of K, n, the scoring function, and the patterning of entries in the various sequences. In some cases, such as the Page test just discussed, the normal approximation is easy to demonstrate and is

probably very good if K is reasonably large. In this latter case, the distribution of the index is generated by a sum of independent random variables, and each individual term is itself asymptotically normal under mild regularity conditions. Most ideally, however, any observed index would be so extreme that a simple Chebyshev inequality in conjunction with the available moments would be sufficient to guarantee significance of the index at some adequate level. Alternatively, approximate permutation tests, such as those discussed by Hope (1968), Cliff and Ord (1973), Edgington (1969), and others, could be used. Given the continuing reduction in computer costs, this latter alternative of sampling from the complete distribution may be the most appropriate to follow in the years to come (see Hubert, 1979, for an illustration of how an approximate permutation test could be carried out).

Indices

Although the raw index Λ_K is sufficient for hypothesis testing, researchers will typically desire some normalized version of Λ_K as a final measure of concordance or agreement. Many different normalizations are possible, but several forms appear continually in the literature. For instance, if it is assumed that Λ_K is nonnegative and that large values of Λ_K denote greater degrees of *concordance* (where this term is being used generically), two general expressions can be given as

$$\Lambda_K/\text{maximum } \Lambda_K; \tag{3}$$

$$[\Lambda_K - E(\Lambda_K)]/[\text{maximum } \Lambda_K - E(\Lambda_K)]. \tag{3a}$$

If the "keying" of Λ_K were in the opposite direction and were denoted by Λ'_K, these two indices would be given as

$$1 - \frac{\Lambda'_K}{\text{maximum } \Lambda'_K}; \tag{4}$$

$$1 - \frac{\Lambda'_K}{E(\Lambda'_K)}. \tag{4a}$$

The indices in Expressions 3 and 4 and in Expressions 4 and 6 are analogues, and to transform one expression to the other, the raw index Λ'_K in Expressions 5 and 6 is rewritten as $\max \Lambda_K - \Lambda_K$. Here, Λ_K is the raw index used in Expressions 3 and 4 and we note that since Λ_K is nonnegative, maximum Λ'_K = maximum Λ_K.

The indices in Expressions 3 and 4 lie between 0 and 1; those in Expressions 3a and 4a are bounded above by 1 but may take on negative values as well. An example of Expression 4 would be Kendall's coefficient of concordance W based on the index Λ_K defined as K times the sum of squares for interaction. The form given by Expression 3a is Cohen's (1968) general expression for his index of nominal scale response agreement kappa. Finally, the measure in Expression 4a provides the basis for one version of Spearman's rank order correlation coefficient based on Γ_2 as well as for the general degree-1 statistic introduced by Hildebrand, Laing, and Rosenthal (1977) in a related nominal scale response-agreement context, using arbitrary weights on the cells of a contingency table.[2]

Although a host of different normalized indices are possible (cf. Hubert & Levin, 1976), the forms given by Expressions 3a and 4a appear to be of continuing importance in the behavioral sciences because they are "corrected for chance." It is important to remember, however, that any normalization is more or less arbitrary, and hypothesis testing can be carried out using only the raw index Λ_K.

Discussion

There are a number of directions in which the preceding material can be extended. For example, instead of permuting the entries within the rows of the K by n table, suppose the elements can be rearranged throughout the table. This inference model has been discussed in the context of nominal scale response agreement (Hubert, in press), and it has been pointed out that when K is 2, the index Γ_2 provides an unnormalized intraclass correlation.

In addition to alternative inference models, other applications of the sequence-comparison

---

[2] In many cases it is traditional to define maximum Λ_K or maximum Λ'_K from some ideal case, for example, untied ranks, and to treat this quantity as a constant irrespective of the configuration of ties in the K by n table.

notion could be developed. For instance, Cochran's Q statistic and McNemar's test for correlated proportions are really special cases of Friedman's test (see Lehmann, 1975), and thus, these techniques could be rephrased within the K sequence-comparison framework. Also, a variant on the type of degree-1 confirmatory analysis developed by Hildebrand, Laing, and Rosenthal (1977) for a bivariate table could be extended to a K-way contingency table by using the sequence-comparison interpretation relevant to nominal scale response agreement.

What is important in all of these applications or extensions is the recognition of problem commonality. The field of nonparametric statistics is very broad and diversified, and consequently, general organizing principles may be of immense pedagogical help to a student attempting to organize the field into a coherent cognitive structure.

References

Cliff, A. D., & Ord, J. K. *Spatial autocorrelation.* London: Pion, 1973.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, *70*, 213–220.

David, F. N., & Barton, D. E. *Combinatorial chance.* New York: Hafner, 1962.

Edgington, E. S. *Statistical inference: The distribution-free approach.* New York: McGraw-Hill, 1969.

Fliess, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323–327.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 1937, *32*, 675–701.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. *Prediction analysis of cross-classifications.* New York: Wiley, 1977.

Hoeffding, W. A combinatorial central limit theorem. *Annals of Mathematical Statistics*, 1951, *22*, 558–566.

Hope, A. C. A. A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B*, 1968, *30*, 582–598.

Hubert, L. Kappa revisited. *Psychological Bulletin*, 1977, *84*, 289–297.

Hubert, L. A general formula for the variance of Cohen's weighted kappa. *Psychological Bulletin*, 1978, *85*, 183–184.

Hubert, L. J. Matching models in the analysis of cross-classifications. *Psychometrika*, 1979, *44*, 21–41.

Hubert, L. Alternative inference models based on matching for a weighted index of nominal scale response agreement. *Quality and Quantity*, in press.

Hubert, L., & Levin, J. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 1976, *83*, 1072–1080.

Kendall, M. G. *Rank correlation methods* (4th ed.). London: Griffin, 1970.

Lehmann, E. L. *Nonparametrics.* San Francisco: Holden-Day, 1975.

Page, E. B. Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 1963, *58*, 216–230.

Pirie, W. R., & Hollander, M. A distribution-free normal scores test for ordered alternatives in the randomized block design. *Journal of the American Statistical Association*, 1972, *67*, 855–857.

Puri, M. L., & Sen, P. K. *Nonparametric methods in multivariate analysis.* New York: Wiley, 1971.

Spearman, C. "Footrule" for measuring correlation. *British Journal of Psychology*, 1906, *2*, 89–108.

Received May 22, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1107–1131

Choosing Between Predictable and Unpredictable Shock Conditions: Data and Theory

Pietro Badia
Bowling Green State University

John Harsh
State University of New York at Geneseo

Bruce Abbott
Bowling Green State University

This article reviews the literature on predictability and describes the factors that affect choice. Particular emphasis is given to the reliability of basic findings, including replications and failures to replicate. Behavioral measures related to choice are also reviewed, and some physiological correlates of predictable and unpredictable shock are noted. The data allow several firm conclusions to be drawn regarding preference, notably that (a) rats (albino, hooded, male, female) prefer predictable shock conditions; (b) they prefer predictable conditions whether shock is avoidable, escapable, or inescapable, and whether it is scrambled or unscrambled grid shock; (c) this preference occurs with different procedures, apparatus, and shock delivery systems, such as water electrodes or electrodes attached to the tail, back, ears, or pubis bone; (d) fish and birds also prefer the signaled condition; and (e) although the preference is robust, it is affected by shock intensity, signal duration, intershock intervals, amount of training, and the dependability of shock-free periods. Other factors that may affect preference are also noted. Finally, the theoretical views of conditioned reinforcement, of information, of preparation, and of safety are evaluated, and their strengths and weaknesses are described.

Research over the last two decades reveals clearly that preference for schedules of both positive reinforcers (e.g., food) and negative reinforcers (e.g., shock) is governed by more than reinforcer-related variables such as rate, magnitude, and duration. Subjects confronted with a choice between a predictable versus an unpredictable shock or food condition most often choose the predictable one, even though all other factors are constant.[1] This result is perhaps not surprising, since predictability may afford the subject the opportunity to prepare for the reinforcer in a way that minimizes its aversiveness or maximizes its attractiveness. We argue, however, that despite its intuitive appeal, this "preparation" hypothesis does not adequately account for current data. We also examine the strengths and weaknesses of alternative explanations.

In our discussion, we focus on predictable and unpredictable shock conditions with nonhuman subjects. We first review studies in which animals are given a choice between

This research was supported in part by Grant GB 33725 from the National Science Foundation, the Faculty Research Committee of Bowling Green State University, and the Research Foundation of the State University of New York.

Bruce Abbott is now at Indiana University—Purdue University at Fort Wayne.

Requests for reprints should be sent to Pietro Badia, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.

[1] There are times in the appetitive situation when subjects choose an unpredictable condition over a predictable one (e.g., Herrnstein, 1964). In Herrnstein's study and in others, animals consistently preferred a variable interval schedule for food over a fixed interval schedule for food of equal value. Such findings suggest that predictability based on periodicity and predictability based on signaling may have different properties or that aperiodicity may have effects other than those produced by lack of signaling.

predictable and unpredictable shock; we discuss the basic phenomenon and some factors that affect the strength of preference. Then we review some of the issues concerning basic findings, research methods, replications, and failures to replicate. In a separate section, we note some behavioral and physiological correlates of predictable and unpredictable shock. Finally, we assess the different theoretical interpretations that have appeared in the literature on choice and predictability. Specifically, we assess the preparation hypothesis, the discriminative stimulus ($\overline{\mathrm{CS}}$) or safety hypothesis, the information hypothesis, and the conditioned reinforcement hypothesis.

Initial Data and Theory

Earlier theorizing about the factors controlling behavior in aversive situations originated in studies focusing on the properties acquired by stimuli preceding shock. One commonly held view is that the pairing of previously neutral stimuli with an aversive stimulus, such as shock, results in conditioned aversiveness (e.g., Mowrer, 1947; Schoenfeld, 1950). This view, which appeals to a conditioned reinforcement process, has played a particularly important role in theoretical accounts of avoidance, escape, and punishment behavior. With respect to choosing between signaled and unsignaled shock, this view predicts that schedules of unsignaled shock will be less aversive than (preferred over) schedules of signaled shock because the latter include the aversive properties of shock plus those of the signal.

Coppock (1954) reported one of the first studies related to preference and to the properties acquired by stimuli preceding aversive events. During training, rats were exposed to either signaled (cue preceded shock) or unsignaled (cue occurred during shock) tail shock while in a restraining apparatus. Shocks were omitted during a testing phase, but by turning its head to one side, the subject could produce the cue previously associated with shock. The analysis of head movement toward the signal suggested to Coppock that stimuli preceding shock acquire greater control over responding than do stimuli occurring during shock. According to Coppock, these findings were unexpected, since the shock signal should have acquired conditioned aversive properties. Coppock concluded that contrary to conditioned reinforcement theory, stimuli preceding shock can acquire positive reinforcing properties but he gave no rationale for how this occurred.

A few years later, Knapp, Kause, and Perkins (1959) gave rats a choice between immediate and delayed shock (Experiment 1), using a T maze. Their subjects preferred the immediate condition. Knapp et al. argued that their findings could not be accounted for by existing conditioned-reinforcement theory, which in this case would predict that painful stimuli immediately following a response should become more aversive than the same stimuli when delayed. In Experiment 2, they gave subjects a choice between delayed shock preceded by a signal and delayed shock followed by a signal. The subjects preferred the signal–shock condition. Arguing again that conditioned reinforcement theory was inadequate to explain their findings, Knapp et al. suggested the preparatory-response hypothesis developed earlier by Perkins (1955). In brief, this hypothesis states that signals preceding appetitive or aversive events allow the organism to prepare for the receipt of stimulation. Such preparation allegedly maximizes the reinforcing properties of appetitive events or minimizes the aversive properties of painful events.

About a year later, Mowrer (1960) described two studies conducted by Mohammed Akhtar. In both studies, signaled or unsignaled shocks were delivered to a grid floor at irregular intervals. If the subjects faced in one direction, they received signaled shocks; if they faced in the other direction, shocks were unsignaled. Shocks were avoidable or escapable. Four of five rats in the first experiment preferred the signaled avoidable shock, and all four subjects in the second experiment preferred the signaled shock condition. Mowrer's interpretation of these results anticipated much of the theorizing that later developed and that is now referred to as the safety hypothesis. In de-

scribing the characteristics of the signal condition, Mowrer stated that a subject "could sharply discriminate between brief periods when it was in real danger (i.e., when the tone was on) and the rest of the time when the rat was perfectly safe and could well afford to 'relax'" (p. 194). In answer to why the rats might seek out a warning signal preceding aversive shock, he stated: "They did not seek the warning signal as such; instead, they sought the situation in which the warning signal occurred, because . . . they experienced less total fear here than in the no-signal situation" (p. 196). Similar interpretations were subsequently offered by others (Badia, Culbertson, & Lewis, 1971; Denny, 1971; Lockard, 1963; Seligman, 1968). The development of various theoretical positions is discussed in detail later.

A theoretical view stressing the importance of information emerged at about the same time that Mowrer's views were made known (Berlyne, 1960). In brief, Berlyne suggested that uncertainty about the occurrence of biologically significant events creates a state of conflict and that stimuli that reduce this conflict (i.e., provide information) are rewarding. This "uncertainty reduction" view clearly predicts that preshock stimuli should be sought out for their informational value. Therefore, the data that were compatible with the views of Coppock (1954), Mowrer (1960), and Perkins (1955) were also compatible with those of Berlyne (1960).

The stage was set for assessing how well the various hypotheses of conditioned reinforcement, preparation, uncertainty, and safety could account for the accumulating data on preference for predictable versus unpredictable shock. Inevitably, subsequent tests involved different procedures and different parameter values, raising both methodological and empirical issues. Some of these remain unresolved, and heated arguments have appeared in the literature (Badia & Harsh, 1977a, 1977b; Biederman & Furedy, 1976b; Harsh, 1978).

What does the literature reveal regarding the phenomenon of choice? Is the phenomenon reliable? Is it robust? What are the boundary conditions? What theoretical interpretations are most favored?

Literature Review

How Reliable Is the Phenomenon?

Since the early studies of Coppock (1954), Knapp et al. (1959), and Akhtar (in Mowrer, 1960), numerous studies have confirmed the preference for signaled over unsignaled shock schedules. A reliable and robust preference for signaled shock has appeared with a variety of choice procedures, shock delivery methods, and species.

Some of the earliest research specifically directed toward the question of preference concerning signaled or unsignaled shock was performed by Perkins, Levis, and Seymann (1963) and by Lockard (1963). Perkins et al. ran rats and found that 14 of the 16 subjects spent about 70% to 80% of the time on the signaled shock side of the shuttle box. When the conditions were reversed, however, a reversal in preference did not occur. Perkins et al. attributed this failure to reverse to the fact that there were only three test days. Reversal was not a problem in a subsequent study by Perkins, Seymann, Levis, and Spencer (1966). Lockard's (1963) results were similar to those of Perkins et al. After 12 sessions, her subjects were spending about 90% of the trials on the signaled side of the apparatus. Control subjects presented with a random signal and shocks were indifferent. Lockard demonstrated that neither signals alone nor random signals and shocks were able to maintain a preference for the signaled condition; a strong preference emerged only when signals preceded shock. Both Perkins et al. and Lockard used unscrambled shock.[2]

Lockard (1965) continued her earlier in-

[2] With unscrambled grid shock, the charge on each grid bar is fixed and alternate grid bars carry opposite charges (positive or negative). With scrambled shock, the charge on each grid bar is rapidly alternated so that in each cycle it is momentarily the opposite of the charge on other bars. Scrambling the shock is designed to eliminate unauthorized avoidance responses (e.g., standing on grid bars of the same polarity) that sometimes occur with unscrambled shock.

vestigation of preference using two different shock (unscrambled) intensities (.221 mA or .236 mA) and four signal–shock intervals (5 sec, .5 sec, 0 sec, or random). Four groups had the same level of shock for both compartments, and four groups had a 7% higher shock level for the signaled side. Only subjects receiving equally intense shocks on both sides in the .5-sec or 5-sec groups tended to prefer the signaled condition. Subjects with the 0-sec signal–shock interval performed the same as the random-interval group did. Also, the level of preference was far weaker than it had been in the first study (Lockard, 1963). One reason why Lockard failed to replicate the high preference found in her earlier work may be that the shock levels were too low in the later study. Other variables were examined by Perkins et al. (1966). They used unscrambled shock with a variety of intershock intervals, shock durations, and signal durations. Only a brief summary of their extensive findings can be given here. In particular, they showed that parameter values were important determiners of choice for the signaled condition. The most marked preference occurred with 12 to 60 shocks per hour, an 18-sec signal duration, and either a .5-sec or 5-sec shock. When the signal duration was decreased, preference was weaker. Preference was also weaker when the intershock interval resulted in 2 shocks per hour. An interesting and important part of their study was the delivery of shock through ear clips for one group (Experiment 5). The results of this experiment will be described later in the section on shock delivery systems.

A replication of Perkins et al. (1966) was performed by Furedy and Walters (Note 1) but with an added condition. Half of their subjects received scrambled shock, the other half, unscrambled shock. Subjects receiving scrambled shock spent a higher percentage of time on the signaled side than did subjects receiving unscrambled shock (80% vs. 65%). However, because of equipment failures and because the typical learning curve for preference was not found, the authors doubted the validity of their findings. They performed a second experiment with improved equipment and found that subjects given either scrambled or unscrambled shock preferred the signaled condition. However, the preference was stronger with unscrambled shock. A clear learning curve was also observed over the first 2 days similar to that reported by other researchers. The data of Furedy and Walters suggested that scrambling may have some effect but that it clearly does not eliminate preference for the signaled condition.

Several other studies using a shuttle box and scrambled shock reported that subjects preferred a signaled shock condition over an unsignaled one, and preference was marked for most conditions (Frankel & Vom Saal, 1976; Gliner, 1972; Hymowitz, 1973). Gliner's study also assessed the effects of shock intensity on choice and the effects of signaling shock on the physiology of the organism. He found that a marked preference for the signaled condition developed for both high- and low-shock-intensity subjects, but the preference developed more slowly for subjects receiving high-intensity shock. Also, he found less physiological deterioration when shock was signaled, a phenomenon we describe in more detail later.

Observing Response Procedure and Choice

Except for the early studies of Coppock (1954), Knapp et al. (1959), and Akhtar (in Mowrer, 1960), all studies reviewed thus far used a shuttle box procedure to assess choice, and the subjects preferred the signaled condition. Other procedures have been used with similar results. An extensive series of experiments has been performed using an operant changeover procedure (e.g., Badia & Culbertson, 1972). With this procedure, baseline responding on a changeover lever is recorded while a subject is being exposed (training) to the different conditions of the experiment. Shocks are usually presented noncontingently on some schedule, and different correlated stimuli identify the signaled and unsignaled conditions. Following training, subjects are given an opportunity to choose the condition in which they prefer to remain (testing). Subjects are placed in an imposed condition of either signaled or unsignaled shock, and a lever press (changeover re-

sponse) changes the condition to the opposite one for a fixed period of time (e.g., 1 min.). At the end of this fixed period, the initial imposed condition is then reinstated and remains in effect until another changeover response occurs. This procedure is similar to the one introduced by Wyckoff (1952).

The first preference studies using the changeover procedure were reported by Badia, Culbertson, and Lewis (1971) using signaled and unsignaled avoidable shock, and by Badia and Culbertson (1972) using either signaled or unsignaled escapable (Experiment 1) or inescapable (Experiment 2) shock. Subjects often spent 90% or more of the time in the signaled condition, and it did not matter whether shock was avoidable, escapable, or inescapable. The strong preference for the signaled condition was shown within each subject repeatedly over a series of acquisition and extinction conditions. Three different extinction conditions were also given to identify stimuli controlling the preference. Under extinction conditions, a response produced (a) the stimulus (signal) identifying the shock period, (b) the stimulus identifying the shock-free period, or (c) neither stimulus. Changeover responding was greatest under the second condition, intermediate under the first, and least under the third. An exact replication of Badia and Culbertson's (1972) study was undertaken by Lewis and Gardner (1977), and the data obtained were virtually identical.

Another series of experiments assessed the robustness of the phenomenon (Badia, Coker, & Harsh, 1973; Badia, Culbertson, & Harsh, 1973). Subjects in each experiment were initially given a choice between signaled and unsignaled shock schedules, with shock parameters equal in the two conditions. Then density, duration, or intensity of signaled shock was systematically increased over that of unsignaled shock. Subjects chose the signaled condition over the unsignaled one when shock parameters were equal and continued doing so when signaled shocks were up to four times more dense, four to nine times longer, or two to three times more intense than unsignaled shocks. A control condition for the shock density study showed that subjects would choose a lower density shock schedule over a higher one when all shocks were unsignaled. However, choice of the lower density schedule was eliminated by signaling the shocks on the higher density schedule.

Most experiments testing the effects of predictable versus unpredictable aversive shock conditions have employed warning signals to make shock and shock-free periods predictable. One can also increase predictability by making the occurrence of shocks temporally regular. Badia, Harsh, and Coker (1975) assessed the relative effectiveness of temporal regularity and signaling on preference using the changeover procedure. Subjects chose between fixed-time (FT) and variable-time (VT) shock schedules under several signaling conditions: (a) unsignaled FT versus unsignaled VT shock, (b) signaled FT versus signaled VT shock, and (c) unsignaled FT versus signaled VT shock. Subjects chose the FT over the VT shock schedule when both were unsignaled and again when both were signaled. They chose the VT over the FT schedule only when the VT shocks were signaled and the FT shocks unsignaled.

Shock Delivery Systems and Choice

Some authors have been critical of the methods used to study signaled and unsignaled shock (Biederman & Furedy, 1976b). One of the major criticisms is that preference for the signaled condition occurs only when shock is modifiable, that is, only when subjects are able by means of postural adjustment on the shock grid to partially or totally avoid the shock or to minimize its aversiveness. The criticism is of obvious import to both theoretical and methodological issues. If true, it would bear directly on the development of theoretical statements. A number of studies relate directly to the alleged problem of differential shock modification. Various procedures have been used to eliminate the problem or to monitor the extent to which it might occur. To ensure that subjects do not differentially modify the shock, some investigators have delivered shock through ear clips (Perkins et al., 1966), through tail electrodes (Coppock, 1954; Miller, Daniel, & Berk, 1974; Miller, Marlin, & Berk, 1977),

through implanted electrodes (Cotsonas, 1972; Griffin, Honaker, Jones, & Pynes, 1974), or through water electrodes (Fisher & Badia, 1975). The use of attached electrodes makes it virtually impossible to modify shock through unauthorized escape or avoidance responses. Other researchers have monitored and measured the amount of shock received by subjects when shock was signaled and when it was unsignaled (Lockard, 1963). A consistent pattern emerges from this research.

The first experiment using surface electrodes, in this case ear clips, was Experiment 5 of the Perkins et al. (1966) study. These investigators found that 13 of the 16 subjects preferred the signaled condition on Day 1. Of the 10 subjects completing the second day of testing with the ear clips still attached, 5 preferred the signaled condition. For the 3 subjects completing 3 days and 2 subjects completing 4 days of testing, the preference for the signal condition was marked. Clearly, the subjects of Perkins et al. preferred signaled over unsignaled shock conditions even with shock applied directly to the body. Other studies using tail shock have been reported by Miller et al. (1974) and Miller et al. (1977). The subjects in these studies also chose the signaled over the unsignaled condition and in some cases over successive reversals.

Findings similar to those for tail shock have been reported with back electrodes by Cotsonas (1972). Brass safety pins were inserted into the lower back in the first part of the experiment. Two subjects were tested, and both preferred the signaled condition; only one later showed a reversal when the conditions were changed. In the second part of the experiment, eight subjects were run. Seven of the eight preferred the signaled condition; all seven succeeded in three reversal conditions. The next experiment kept subjects in the same condition for 15 days prior to reversing for 15 days. The subjects did not continue to remain on the signaled side. The latter outcome is possibly due to problems arising from chronically implanted electrodes. Other investigators using grid shock have not found any attenuation of preference over time.

The generality of the preference for signaled over unsignaled shock in terms of shock delivery systems and in terms of species was increased by Griffin et al. (1974) and by Fisher and Badia (1975). Griffin et al. implanted electrodes in pigeons by attaching them to the pubis bone and gave their subjects a choice between the two shock conditions. Two pigeons were run using a procedure similar to that of Badia and Culbertson (1972). Variability tended to be high; nevertheless, one subject clearly changed to the signaled condition, and the other showed a similar but less marked trend under the high-intensity shock condition. Fisher and Badia (1975) used the changeover procedure with goldfish in a shuttle box. Water electrodes distributed shock evenly throughout the shuttle box. The subjects preferred the signal condition, and their performance was similar to that of rats tested under similar conditions, including performance under the three different extinction conditions used by Badia and Culbertson (1972).

Others have used different procedures to determine if signaled shock was more modifiable than unsignaled shock. These studies monitored the development of skeletal preparatory responses to the signal (Badia & Abbott, in press; Biederman & Furedy, 1973, Experiment 1; Biederman & Furedy, 1976b; Furedy & Biederman, 1976; Lockard, 1963; Marlin, Berk, & Miller, 1978; Abbott & Badia, Note 2). Lockard used a pen recorder to monitor when subjects avoided unscrambled shock. Only about 1.4% of the shocks were avoided; of these, equal numbers were avoided by the signaled shock group and random signal group. Lockard also failed to detect any systematic differences in the postural behavior of the two groups.

Different results were obtained in a series of experiments by Biederman and Furedy in which the current flow through the subject was monitored during shock. Biederman and Furedy (1973, Experiment 1) found that the degree of unscrambled shock attenuation correlated weakly (.40) but significantly with the degree of preference for the signaled condition. Subjects receiving scrambled grid shock or tail shock showed no attenuation and no preference. Furedy and Biederman

(1976, Experiment 3) obtained similar results using Lockard's (1963) shuttle box procedure with the training phase omitted. Unscrambled shock was used. Modification of shock was again observed, and there was a relation between degree of modification and degree of preference for the signaled condition.

The reports of Biederman and Furedy seem to provide evidence that modification of shock can be an important determinant of preference. However, the studies may be faulted. First, the findings provide no direct evidence on differential modification of shock between the signaled and unsignaled conditions. Separate modification scores were not reported. If preference for the signaled condition is related to better avoidance or reduction of shock in the signaled condition, this needs to be demonstrated. Second, even if one is willing to accept the observed correlation between preference and overall modification as suggestive of differential modification, there is another problem. Biederman and Furedy used exceptionally long shocks (5 sec), whereas most researchers used much briefer shocks (e.g., .5 sec; see Table 1) specifically to minimize the problem of avoidance. Attenuation, if it does occur, is more likely with long rather than short shocks. It may be that shock attenuation with long durations is a sufficient but not a necessary condition for preference to emerge. Another problem relates to training, that is, the exposure of subjects to the signaled and unsignaled conditions prior to their being given a choice. This phase provides subjects the opportunity to associate each condition with its correlated stimulus. Biederman and Furedy obtained their results only after training was eliminated. Other researchers have found that considerable experience with the signaled and unsignaled conditions is necessary before subjects acquire a preference for the signaled condition (e.g., Badia & Culbertson, 1972). The parameters chosen by Biederman and Furedy would appear to minimize a preference for signaled shock based on associative factors and maximize the potential for differential modification of signaled and unsignaled shocks. Therefore, Biederman and Furedy may be studying a different phenomenon. A chronological listing of studies investigating choice under signaled and unsignaled conditions is located in Table 1. The listing includes procedures used, parameters varied, and results obtained.

Evidence suggesting that differential modification of shock may occur was reported by Marlin et al. (1978) and was based on observation of subject postures during signals and during the intertrial interval but not during shock. Following preference testing in which the subjects chose the 100% signal condition over one in which signals preceded shock only 80% of the time, each subject was observed across 20 shocks. The subjects were found to be rearing at shock onset 64% of the time and appeared to avoid 6.5% of the shocks as indicated by the lack of response. The incidence of full rearing (front paws off the grid, body axis more than 45° from horizontal) was only marginally greater during signals than in their absence (30% and 29%, respectively), but partial rearing (body axis less than 45° from horizontal) occurred far more frequently during signals than during signal absence (59% and 3%, respectively). Asserting that rearing responses modify shock, Marlin et al. concluded that scrambled-grid shock allows shock modification and should be replaced with a fixed-electrode preparation to exclude its possibility. Their own fixed-electrode studies (e.g., Miller et al., 1977) convinced them, however, that the preference for signaled shock is not solely determined by modification.

Although the Marlin et al. (1978) results are suggestive, the lack of interrater reliabilities, the failure to use a blind procedure to rate modification to signaled and unsignaled shocks, and the failure to document a relation between observed postures on the grid and actual modification of shock reduce the value of these data. A study that is not open to these criticisms was conducted by Badia and Abbott (in press). The changeover procedure was used to assess preference for signaled over unsignaled shock, and all parameters were similar to previous studies (e.g., Badia & Culbertson, 1972). Additional equipment permitted the measurement of current flow through the subject. The study revealed

(*text continued on p. 1117*)

Table 1
Studies Involving Choice Between Signaled and Unsignaled Shock Schedules Reported Between 1959 and 1978

Investigator	Choice procedure	Parameter	Test session	Results
Knapp, Kause, & Perkins (1959; Experiment 2)	T maze	Unscrambled, .7-sec, 60-V shock; 30-sec signal	50 (8 trials/day)	80–90% of free-choice trials to signal side after 21 sessions
Akhtar (cited in Mowrer 1960)	Orientation in revolving cage	Grill floor, VT 90-sec shock (escapable and avoidable); signal duration not specified	Multiple approximately 25-min sessions	4 of 5 subjects chose signaled schedule
Lockard (1963)	Shuttle box	Unscrambled, 2-sec, .28-mA, VT 1.9-min shock; 5-sec signal	12 1-hour sessions	90% of trials on signaled side after 12 sessions
Perkins, Levis, & Seymann (1963)	Shuttle box	Unscrambled, .5-sec, 500-V, VT 5-min shock; 3-sec signal	6 10-hour sessions	75–80% of time on signaled side; unsuccessful reversal
Lockard (1965)	Shuttle box	Unscrambled, VT 1.9-min. 221-mA or .236-mA shock (duration not specified); signal duration varied	12 1-hour sessions	Preference related to conditions; weak preference at best
Perkins, Seymann, Levis, & Spencer (1966)	Shuttle box	Unscrambled or direct, 500-V shock, duration, ITI, and distribution varied; signal duration varied	6–9 10-hour sessions	Up to 90% of time on signaled side; preference related to conditions; successful reversals
Furedy & Walters (Note 1; Experiment 2)	Shuttle box	Scrambled and unscrambled, 5-sec, mA, VT 2-min shock; 3-sec signal		Preference for signaled side stronger with unscrambled shock
Badia, Culbertson, & Lewis (1971)	Changeover	Scrambled, .75-mW or 1-mA, .32-sec avoidable shock; 5-sec signal	Multiple 6-hour sessions	All subjects changed to signaled schedule
Badia & Culbertson (1972)	Changeover	Scrambled, .75-mW, escapable (Experiment 1) or inescapable (Experiment 2), .5-sec, VT 2-min shock; 5-sec signal	Multiple 6-hour sessions	All subjects changed to signaled schedule
Gliner (1972)	Shuttle box	Scrambled, 2-sec, VT 2.5-min shock (intensity varied); 10-sec signal	5 6-hour sessions	Signaled shock preferred by both high- and low-intensity-shock groups
French, Palestino, & Leeb (1972)	Shuttle box	Unscrambled, <1-mA, 1-sec, VT 90-sec shock; signal duration varied	Multiple 1-hour sessions	Preference for signaled shock by all groups
Cotsonas (1972; Experiments 2 & 3)	Shuttle box	Direct, .5-sec, 2.8-mA, VT 1-min shock; 3-sec signal	Multiple 90-min sessions	Preference for signaled schedule evident, some difficulty with reversals

Table 1 (*continued*)

Investigator	Choice procedure	Parameter	Test session	Results
Badia, Coker, & Harsh (1973)	Changeover	Scrambled, .5-sec, 75-mW, VT schedule varied; 5-sec signal	Multiple 6-hour sessions	Preference for signaled schedule even when signaled shocks 2–8 times more frequent
Badia, Culbertson, & Harsh (1973)	Changeover	Scrambled, VT 2-min shock duration and intensity varied); 5-sec signal	Multiple 6-hour sessions	Preference for signaled schedule even when signaled shocks were up to 9 times longer and 6 times stronger
Hymowitz (1973)	Shuttle box	Scrambled, .5-sec, .4-mA, VT 65-sec shock; 5-sec signal	10 30-min sessions	Up to 80% of time on signaled side after 4 sessions
Biederman & Furedy (1973)	Concurrent chain schedule	Unscrambled, scrambled, or tail-shock, 5-sec, 1.0-mA, VT 45-sec shock; 5-sec signal	1 approximately 4-hour session	Preference for signals with unscrambled shock only
Miller, Daniel, & Berk (1974)	Shuttle box	Direct, .5-sec, .4-mA, VT 2-min shock; 10-sec signal	30–50 3-hour sessions	Clear preference for signaled schedule with reversals
Harsh & Badia (1974)	Changeover	Scrambled, .5-sec, 2-mA, VT 12-min shock; 60-sec signal	Multiple 3-hour sessions	Preference for signaled schedule
Griffin, Honaker, Jones, & Pynes (1974)	Changeover	Direct, .5-sec, 60–80-V, VT 90-sec shock; 5-sec signal	Multiple 3-hour sessions	Pigeons preferred signaled schedule, although considerable variablity
Fisher & Badia (1975)	Changeover	Direct, .5-sec, 8-V, VT 2-min shock; 5-sec signal	Multiple 90-min sessions	All fish preferred signaled schedule
Arabian & Desiderato (1975)	Shuttle box	Scrambled, .5-sec, 2-mA shock; 5-sec signal	10 sessions	See text
Badia, Harsh, & Coker (1975)	Changeover	Scrambled, .5-sec, .8-mA, FT or VT 1-min shock; 5-sec signal	Multiple 6-hour sessions	Preference for predictable shock schedules
Harsh & Badia (1975)	Changeover	Scrambled, .5-sec, VT 2-min shock; 5-sec signal	Multiple 6-hour sessions	Choice for signaled schedule related to shock intensity
Crabtree & Kruger (1975)	Shuttle box	Scrambled, 2-sec, 110-V, VT 2-min shock; signal duration varied	1 12-hour session	No preference observed
Harsh & Badia (1976)	Changeover	Scrambled, .5-sec, .75-mW shock; 30-sec signal	Multiple 6-hour sessions	Choice related to average ITI
Badia, Harsh, Coker, & Abbott (1976)	Changeover	Scrambled, .5-sec, 75-mW shock; 5-sec signal	Multiple 6-hour sessions	Choice related to $p(US \mid CS)$ and $p(US \mid \overline{CS})$

(*table continued*)

Table 1 (*continued*)

Investigator	Choice procedure	Parameter	Test session	Results
Furedy & Biederman (1976)	Shuttle box	Scrambled or unscrambled, 5-sec, .6-mA, VT 2-min shock; 3-sec signal	2 3-hour sessions	No preference with pretraining; without pretraining preference obtained with unscrambled shock only
Biederman & Furedy (1976a)	Changeover	Scrambled, .5-sec, 75-mW, VT 2-min shock; 5-sec signal	Multiple 6-hour sessions	Only 1 of 16 subjects clearly preferred signaled schedule
Biederman & Furedy (1976c)	Shuttle box	Scrambled, 5-sec, .6-mA, VT 2-min shock; 3-sec signal	50 3-hour sessions	Grouped data indicated no preference
Frankel & Vom Saal (1976)	Shuttle box	Scrambled, 1-sec, .5-mA, VT 2-min shock; 12-sec signal	4 10-hour sessions	Subjects spent greater time under signaled conditions
Abbott & Badia (in press)	Changeover	Scrambled, .5-sec, 1-mA, VT 120-sec shock; signal duration varied	Multiple 6-hour sessions	Preference for the signaled condition varied strongly with signal duration
Abbott & Badia (Note 2; Experiment 3)	Changeover	Scrambled, .5-sec, 75-mW or 1.0-mA, VT 2-min shock; 5-sec signal	Multiple 6-hour sessions	Grid contact time equal with signaled and unsignaled shock; choice for former
Collier (1977)	Shuttle box	Scrambled, .75-sec, 58-V, VT 6.7-min shock; 10-sec signal	5 20-min sessions	See text
Miller, Marlin, & Berk (1977)	Shuttle box	Direct or scrambled, .5-sec, intensity varied, ITI varied shock; 5-sec signal	Multiple 1.5-, 2-, or 4-hour sessions	Preference related to parameters; most subjects strongly preferred the signaled condition
Lewis & Gardner (1977)	Changeover; concurrent bidirectional changeover	Scrambled, .5-sec, 75-mW, VT 120-sec shock; 5-sec signal	Multiple 6-hour sessions	All subjects preferred the signaled condition (changeover), but only experienced subjects preferred the signaled condition in the bidirectional procedure
Safarjan & D'Amato (1978)	Changeover	Scrambled, .5-sec, .8-mA, VT 1-min shock; 5-sec signal	Multiple 4-hour sessions	Preference for the signaled condition related to the dependability and duration of the shock-free period

Note. VT = variable time; ITI = intertrial interval; FT = fixed time; US = unconditioned stimulus; CS = conditioned stimulus. All subjects are rats unless otherwise indicated.

that whether shock was signaled or unsignaled made no difference in the duration of contact with the grid bars. If anything, there was a suggestion that 10 of the 12 subjects actually received slightly longer shock durations when shock was signaled. Also of interest were oscilloscope tracings. These tracings indicated that the subjects were rapidly making and breaking grid contact during shock primarily because of running. Such activity permits reduction in grid-contact time for both signaled and unsignaled conditions, but it does not permit the kind of response that would allow precise control of current flow. We should also note that the Badia and Abbott findings confirm those reported earlier by Lockard (1963).

Conclusion

Based on this portion of the literature review, a number of firm conclusions are warranted. Clearly, organisms prefer signaled over unsignaled shock conditions, whether shock is avoidable, escapable, or inescapable. Preference for the signaled condition has occurred in rats, in pigeons, and in fish. When the conditions of the experiment are reversed, preference has also reversed, although an occasional failure to reverse with reversed conditions has been noted.

With rats, a preference for the signaled condition has been found with both males and females, and with hooded and albino animals. The same results have been reported with different apparatus and different procedures. A preference for the signaled condition develops whether shock is scrambled or unscrambled and with various shock delivery systems or surface electrodes, such as electrodes attached to the tail, to the back, to the ears, or to the pubis bone, or with water electrodes. In addition, studies measuring the duration of scrambled shock received by subjects under signaled and unsignaled conditions suggest that differential shock modification, shock avoidance, and decreased shock duration are not necessary conditions for preference. Finally, we can conclude that the preference is robust in that subjects prefer longer, stronger, and more dense signaled shock over shorter, weaker, and less dense unsignaled shock.

Factors Affecting Preference

There seems to be little question that subjects prefer signaled shock conditions and for reasons other than those related to shock modification. We now discuss the factors that strengthen or weaken preferences.

One factor affecting the strength of preference is shock intensity. Harsh and Badia (1975) used constant-current scrambled shock ranging in intensity from .15 mA to 1.0 mA, in steps of either .15 mA or .20 mA. They found that the amount of time spent in the signaled condition varied systematically with shock intensity over the lower and middle range of intensities used. Subjects did not choose the signaled condition at low shock intensities.

Another factor demonstrated to affect choice is the average intershock interval (Harsh & Badia, 1976). These investigators used a constant-wattage scrambled shock of .5 sec over six variable-time intershock intervals and a constant 30-sec signal. The intershock intervals averaged 510 sec, 270 sec, 150 sec, 90 sec, 60 sec, and 45 sec. Choice of the signaled condition was directly related to the average intershock interval for six of the eight subjects in that short intershock intervals weakened preference, and long intershock intervals strengthened it. However, since only one signal duration was used, it cannot be determined from this study whether it was the absolute shock-free period or the ratio of the shock period (signal present) to the shock-free period (signal absent under the signaled condition) that controlled choosing the signaled condition.

The question of whether preference for the signaled condition is controlled by stimuli identifying the shock period (signal) or the shock-free period (signal absence) is important empirically and theoretically. One way of evaluating the role that these stimuli assume is by varying their dependability. In a study by Badia, Harsh, Coker, and Abbott (1976), the dependability of the signal in identifying a shock period was varied by

holding the total number of signals constant at 180 and varying parametrically the number of shocks from 180 to 3. Under these conditions, the probability of shock in the event of a signal varied, that is, p (US|CS) $\neq 1.0$, but the probability of shock in the absence of the signal always remained constant, that is, p (US|$\overline{\text{CS}}$) = 0. Badia et al. found that subjects changed to the signaled condition when the value of p (US|CS) = 1.0 and also when the probability was systematically reduced to less than 1.0. Apparently, the dependability of the signal that identified a shock period was not important as long as the probability of safety was not degraded, that is, as long as p (US|$\overline{\text{CS}}$) remained at 0. A variety of additional conditions, including controls for the intershock interval and sensory stimulation, were run.

In the second experiment, Badia et al. (1976) degraded the dependability of the stimulus identifying the shock-free period (safety), that is, p (US|$\overline{\text{CS}}$) $\neq 0$, while keeping constant the dependability of the stimulus identifying the shock period, that is, p (US|CS) = 1.0. The results showed that as the dependability of safety varied, preference for the signal condition also varied. When safety was dependably identified, preference for the signaled condition was strong, but when safety was undependably identified, preference for the signaled condition weakened. These results suggest that stimuli identifying shock-free periods are more important than stimuli identifying shock periods. The results also raise some interesting questions concerning the relative value of safety. Since subjects chose the signaled condition when safety was a nonzero value, that is, p (US|$\overline{\text{CS}}$) $\neq 0$, it is obvious that safety need not be absolute. Results similar to those of Badia et al. have been reported by Safarjan and D'Amato (1978). These investigators also concluded that preference for the signal condition was strongly related to the safety function of signal absence but not to the warning function of signal presence.

We have already noted the importance of signal duration in the early studies of preference (Perkins et al., 1966). Similar findings were reported by French, Palestino, and Leeb (1972). These early findings have been confirmed and supplemented by other research (Abbott & Badia, in press) using short signal durations with a different procedure and in a different apparatus. It was found that subjects generally did not prefer the signaled condition when the signal duration was less than 1 sec. Signal durations of 1.5 sec or longer, however, resulted in a strong preference. As noted, Perkins et al. found that preference was strongest with signal durations of 18 sec; French et al. found preference strongest with 30-sec durations. The theoretical implication of these findings is discussed later.

Other studies are relevant to an analysis of the factors affecting preference, and several investigators have assessed the attractive or aversive properties of stimuli that serve warning and safety functions (Arabian & Desiderato, 1975; Collier, 1977; Harsh & Badia, 1974). Harsh and Badia gave rats a choice between signaled and unsignaled shock while they were responding on a variable-interval food schedule. All the subjects chose the signaled condition, even though responding for food in the presence of the signal was suppressed. The rate of responding for food was lowest in the presence of the signal and highest in its absence. An intermediate rate of responding occurred under the unsignaled shock condition. Although it is clear that response suppression cannot always be used as an index of aversiveness (cf. Rachlin & Herrnstein, 1969), under these conditions we believe that the results suggest that the presence of the signal is the most aversive, that stimuli associated with the unsignaled condition are less aversive, and that the absence of the signal under the signaled condition is the least aversive.

The studies by Arabian and Desiderato (1975) and Collier (1977) assessed the characteristics of stimuli associated with shock periods (danger), stimuli associated with shock-free periods (safety), and signals preceding shock (warning stimuli). Three groups of rats were tested in each of the studies. One group was exposed to a situation in which stimuli identified shock and shock-free periods. In addition, during the shock periods

this group had a signal preceding each shock. Therefore, this group had stimuli identifying safe periods and danger periods, and also warning signals during the danger periods. A second group was essentially the same as the first, but instead of a signal preceding each shock during the shock periods, a random signal was used. For the second group, then, safe and danger periods were identified, but the warning signals were absent. The third group did not have stimuli that identified safe, danger, or warning periods. After training under the above conditions, the subjects were given a choice. Both studies reported that their subjects preferred the condition in which shock and shock-free periods were identified over the condition in which they were not. The subjects also preferred the condition of discriminable shock and shock-free periods when warning signals were added over the condition without discriminable shock and shock-free periods. Different findings were reported, however, when the subjects were given a choice between a condition in which stimuli identified shock and shock-free periods and one in which discriminable periods plus warning signals were present. Collier found that the subjects preferred the condition containing warning signals, whereas Arabian and Desiderato did not.

Conclusion

Our review of this portion of the literature dealing with the factors affecting choice permits a number of conclusions. The literature suggests that choice of the signaled condition does not occur at low levels of shock intensity. Generally, as shock intensity increases, so does choice of the signaled condition. The relation found between choice and shock intensity is similar to that found between performance and shock intensity in avoidable and escapable situations. The dependability of stimuli identifying a shock period appears to be relatively unimportant in terms of choice. Subjects chose the signaled condition even though the dependability varied markedly. On the other hand, the dependability of a stimulus identifying a shock-free or safe period is important. When this stimulus is made relatively undependable, choice of the signaled condition decreases. Long signal durations tend to be more effective than short signal durations. Signal durations of less than 1.5 sec generally do not result in a preference for signaled shock. Finally, very short (45 sec) or very long (120 min) intershock intervals attentuate preference for the signaled condition.

Failures to Replicate

Several studies have obtained preference for a signaled shock condition only when the shock was unscrambled (Biederman & Furedy, 1973, 1976a, 1976c; Furedy & Biederman, 1976). These data suggest that preferences for a signaled condition emerges only when subjects can overtly modify the shock through skeletal-muscular responses. This conclusion clearly disagrees with the findings reviewed earlier showing that subjects prefer a signaled condition with scrambled grid shock, with tail shock, and with water electrodes (e.g., Badia & Culbertson, 1972; Fisher & Badia, 1975; Miller et al., 1974).

Other studies by Biederman and Furedy (1973) have also failed to obtain a preference for signaled shock when it was scrambled. Yet, other investigators using a similar procedure have not had difficulty (Abbott & Badia, Note 2). Various possible reasons for the unusual findings of Furedy and Biederman are described in some detail in Badia and Harsh (1977a, 1977b). In part, they relate to the parameter values chosen by Furedy and Biederman, such as signal duration (3 sec), amount of training (none), length of testing phase (two 3-hour sessions), and intershock interval (45 sec). Some of these values have been shown to be less than optimal for demonstrating preference. In addition, the uncommonly long shocks used (5 sec) may have provided unusual opportunities for the unscrambled-shock subjects to learn competing responses.

We are aware of only one other study that failed to obtain a preference for a signaled shock condition over an unsignaled one (Crabtree & Kruger, 1975). In that study, however, several problems made finding a pref-

erence unlikely. The investigators used only one 13-hour testing session, an *n* of 1 in each cell for a 3 × 5 design, an intershock interval that was partially predictable, and a 3-hour pretest session with only the signal present. The latter procedure may have rendered the signal ineffective through the process of latent inhibition.

Related Findings

Studies using a choice as a dependent measure represent only one portion of a considerably larger amount of literature dealing with the role of predictability in aversive situations. This literature is much too extensive to be reviewed in any detail here; however, the general findings of studies involving measures other than choice are discussed insofar as they relate to the choice literature. (For a recent review dealing with response suppression under signaled and unsignaled conditions, see Hymowitz, 1979.)

Avoidance, Escape, and Punishment

The behavioral significance of signaling aversive events is clearly revealed in studies of avoidance, escape, and punishment. Sidman (1955) published one of the first studies comparing signaled and unsignaled avoidance schedules. Both shocks and signals were under the subject's control in this study. Responses in the presence of the signal postponed the next shock, and responses in its absence postponed both the signal and shock. The subjects allowed the signal to appear rather than postpone it, and overall avoidance responding dropped. Similar findings have been obtained by other investigators (e.g., Badia, Culbertson & Lewis, 1971; Hyman, 1969; Keehn, 1959; Ulrich, Holz, & Azrin, 1964). One explanation of the differences in responding may relate to differences in shock density under signaled and unsignaled schedules. Signals may allow a more effective avoidance strategy (i.e., result in fewer shocks). However, although some investigators have found shock density differences (e.g., Badia, Culbertson, & Lewis, 1971; Ulrich et al., 1964), others have not (e.g., Ayers, Benedict, Glackenmeyer, & Matthews, 1974; Logan & Boice, 1968; Powell, 1976). An explanation of the signaling effect based solely on an increase in response effectiveness is not supported.

Signaling also has a marked effect on behavior under schedules of escapable shock. Badia and Culbertson (1970) compared the behavior of rats under signaled and unsignaled escapable shock schedules and did not find differences in escape latencies. This outcome also suggests that behavioral differences resulting from signaled and unsignaled schedules are not due to response effectiveness. On the other hand, Badia and Culbertson did find clear differences in lever-holding time and in exploration. Holding was less frequent and exploration more frequent under the signaled condition.

Comparisons of signaled and unsignaled punishment schedules have yielded findings similar to those obtained with escape and avoidance (e.g., Church, 1969). Church found that immediate shock resulted in more suppression than did delayed shock. More interesting, however, a signaled punishment group showed less overall suppression than did an unsignaled punishment group, and the distribution of responding was different. Responding in the absence of the signal was higher than in the presence of the signal.

The studies reviewed thus far indicate that signaled and unsignaled avoidance, escape, and punishment procedures have different effects on behavior. When signals are present, they tend to gain control over behavior related to the postponement, termination, and/or prevention of shock. When signals are absent, the prevailing stimuli appear to set the occasion for behaviors not related to the control of shock (e.g., general activity, responding for reinforcement). When signals are unavailable, shock-related behaviors tend to predominate at all times.

Response-Independent Shock and Behavioral Suppression

In contrast to response-dependent shock (Church, 1969), other studies have assessed the effects of response-independent shock on behavior maintained by reinforcement, that is,

conditioned suppression (see Blackman, 1977; Davis, 1968). Some researchers have compared shock schedules with signals to shock schedules without signals (e.g., Brimer & Kamin, 1963; Davis & McIntire, 1969; Davis, Memmott, & Hurwitz, 1976; Holmes, Jackson, & Byrum, 1971; Seligman, 1968; Seligman & Meyer, 1970; Shimoff, Schoenfeld, & Snapper, 1969; Weiss & Strongman, 1969). These studies have shown that when shocks are signaled, the base rate of responding initially drops but gradually recovers over time. When shocks are not signaled, however, the base rate of responding drops to a low level and shows little recovery.

An example of correspondence between variables influencing choice and conditioned suppression concerns the dependability of predictors of shock and shock-free periods. In a study described in more detail earlier, Badia et al. (1976) reported that systematic reduction of the probability with which signals were followed by shock had little effect on preference, whereas degrading the dependability of a stimulus identifying a shock-free period had a marked effect. A similar finding using a conditioned suppression procedure was reported by Nageishi and Imada (1974). They studied the effects of varying the dependability of the shock-free periods on rats' licking behavior. They found that as the dependability decreased, the basal rate of licking also decreased.

Somatic Reactions to Shock

There has been a variety of studies comparing the somatic reactions to signaled and unsignaled shock, and many of the reports indicate that the effects of signaled situations are less severe than the effects of unsignaled ones (Gliner, 1972; Mezinskis, Gliner, & Shemberg, 1971; Price, 1972; Seligman, 1968; Seligman & Meyer, 1970; Simpson, Wilson, DiCara, Jarrett, & Carroll, 1975; Weiss, 1970, 1971a, 1971b, 1971c). The findings by Weiss (1970) are representative. Weiss found marked differences in the stress responses of subjects receiving signaled shocks and of subjects receiving unsignaled shocks. Subjects receiving unsignaled shocks developed more ulcers, lost more weight, and showed higher plasma corticosterone concentrations and higher body temperatures than did subjects receiving signaled shock.

The Weiss (1970) study and others clearly suggest that the somatic consequences of shock schedules are less severe when shocks are signaled. However, results involving somatic measures are not entirely consistent, and apparently contradictory findings have been obtained. Compared to unsignaled schedules, signaled schedules have been associated with heightened adrenal functioning (Bassett, Cairncross, & King, 1973; Paré, 1964), greater weight loss, and higher mortality rates (Brady, Thornton, & DeFisher, 1962; Friedman & Ader, 1965).

An explanation of the conflicting findings related to somatic reactions is not available. Weiss (1977) suggests that an examination of procedural variables reveals some consistencies. That is, somatic reactions to signaled schedules are less severe relative to reactions to unsignaled schedules when a direct-shock delivery system (e.g., tail shock) is used and are usually more severe when grid shock is used. Weiss (1977) related this pattern of findings to rats' coping behavior with different shock delivery systems. According to his view, the inefficient coping attempts associated with grid shock but not with direct shock may lead to more pathological changes. Not all findings are consistent with this notion however (Gliner, 1972; Seligman, 1968; Seligman & Meyer, 1970). It is apparent that additional data are needed to clarify this issue.

Theoretical Views

There are several ways of organizing the literature on predictability and behavior, particularly choice behavior. The findings could be organized along conditioned reinforcement lines. In its most simple form, this view states that neutral stimuli paired with primary reinforcers (e.g., food or shock) acquire reinforcing properties similar to those of the reinforcer. A conditioned reinforcement view can account for the findings obtained in experiments using food that subjects choose situations in which a signal precedes a posi-

tive reinforcer over situations in which no signal is given (e.g., Lutz & Perkins, 1960; Prokasy, 1956). To account for the preference for signaled over unsignaled food, it is assumed that the total amount of reinforcement is greater with the signal than without it, that is, the summation of the conditioned reinforcement occurring to the signal through pairings with food and the reinforcement of the food itself is greater than food reinforcement alone in the unsignaled condition. Generalizing this logic to the aversive situation, analogous reasoning would predict the opposite results, that is, shock plus the acquired aversiveness of the signal should be more aversive than the condition with unsignaled shock alone. As our view indicates, however, subjects under aversive stimulation clearly prefer signaled over unsignaled aversive situations. Obviously, conditioned-reinforcement theory alone cannot account for the literature showing that subjects prefer the signaled-shock condition. Nor can it account for appetitive findings showing that in some cases, subjects prefer unsignaled over signaled appetitive reinforcement (Hershiser & Trapold, 1971).

Another way of organizing the literature on predictability could follow an information-theory and uncertainty-reduction analysis (e.g., Berlyne, 1960). Berlyne's theory is based on the proposition that drive induces uncertainty and conflicts, and on the reinforcing effect of their reduction. All information is considered desirable, and the theory does not provide a basis for selecting information. Rather damaging to the uncertainty reduction view is the fact that it does not predict the findings of Defran (1972), Dinsmoor, Flint, Smith, and Viemeister (1969), Kendall (1973), and Wilton and Clements (1971). The Dinsmoor et al. and Defran studies are most illustrative. These investigators used an observing-response procedure and found that subjects would respond for stimuli identifying food periods (Dinsmoor et al.) or shock-free periods (Defran), but that they would not respond for stimuli identifying shock periods. Also, many investigators have shown that animals prefer information concerning reward over information concerning nonreward (Dinsmoor, Browne, & Lawrence, 1972), even when the information content is equal (e.g., Jenkins & Boakes, 1973; Peterson, Ackil, Frommer, & Hearst, 1972). Other findings difficult for the information–uncertainty-reduction hypothesis to address are those of studies showing that preference is a function of signal duration (e.g., Perkins et al., 1966), of reward magnitude (Mitchell, Perkins, & Perkins, 1965), or of stimulus dependability (Badia et al., 1976). It is apparent that the information hypothesis is not sufficiently developed to apply to much of the current literature, especially the literature on choice.

The two major analyses that have been applied to the choice literature are the preparation hypothesis and the safety hypothesis.

Preparatory Response Hypothesis

According to the preparation hypothesis (Perkins, 1955, 1968), stimuli that precede biologically important events allow subjects to prepare to receive these events. In turn, preparation is thought to minimize the painfulness of aversive stimulation or maximize the attractiveness of appetitive events. These preparatory responses are considered to be classically conditioned responses acquired through the law of effect.

The conditioned reinforcement hypothesis, and the preparation hypothesis predict similar outcomes in appetitive situations but not in aversive situations. Under aversive stimulation, the preparation hypothesis predicts that subjects will prefer signaled over unsignaled shock conditions. According to the earlier theorizing of Perkins (1955), signals preceding shock allow the subject to make preparatory responses (either internal or external) to shock, which reduces its aversiveness. In a later version of the theory, Perkins (1971) generalized preparatory responding to include the entire stimulus situation of shock and shock-free periods. This conception is in sharp contrast to the earlier view (Perkins, 1955), in which the emphasis was placed on specific responses. It is the earlier version of preparation that is most frequently tested and that is most testable.

Advantages of the preparation hypothesis.

One advantage of this view is its parsimony. The preparation hypothesis presents a one-factor view of conditioning in that both instrumental conditioning and classical conditioning are explained in terms of the law of effect. In addition, this view fits nicely with our intuitive notions of the kinds of behavior that should occur in response to signals that predict biologically important events. For example, given the opportunity, subjects should respond in ways that increase the attractiveness or decrease the aversiveness of environmental events. Numerous experiments dealing with escape and avoidance bear out this expectation. Preparatory responses clearly do occur under certain conditions. Another advantage of the preparation view is that it emphasizes the importance of such signal parameters as duration, variability, and dependability in determining choice. As noted, much of the evidence concerning choice under different signal conditions is consistent with preparation. It has been argued that in some situations longer signals allow more adequate preparation than do shorter signals, an outcome that has been found in several studies (Abbott & Badia, in press; French et al., 1972; Perkins et al., 1966). Particularly relevant are the findings of Abbott and Badia showing that signal durations of .5 or 1.0 sec would not support a preference for the signaled shock conditions. According to the preparation hypothesis, these short signal durations simply do not provide adequate preparation time. Signal variability could also be considered important for preparation. To be maximally effective, preparatory responses must be precisely timed, and conditions that allow this would be preferred to conditions that do not. Evidence supporting this view has been found by Safarjan and D'Amato (1977). This study reports that subjects prefer fixed over variable signal durations. The assumption of precisely timed preparatory responses also provides the rationale for predicting preference for immediate over delayed shock (e.g., Knapp et al., 1959).

The data on dependability of the signal reported by Badia et al. (1976) appear incompatible with a preparatory view. However, it may be possible to interpret the data of Badia et al. within a preparatory-response framework by (a) assuming a positive relation between preference and the proportion of shocks to which preparation occurred (Experiment 1) or (b) assuming preparation required little effort so that inappropriate preparation did not affect preference (Experiment 2).

Difficulties with the preparatory hypothesis. The preparatory view can account for a substantial portion of the literature on preference for predictable events. However, the strength of this view is also its weakness in that each successful account requires making a specific assumption about the nature of preparatory responses. Often these assumptions are deduced from the experimental outcomes that they allegedly predict. Since assumptions appropriate to any given outcome can be postulated, it is unlikely that a definitive test of the preparation hypothesis can be made. An important question, therefore, is whether the various assumptions made across experiments represent a unified view of the preparation hypothesis. Unfortunately, when the assumptions are viewed in this larger context, they often conflict. For example, to account for the choice of immediate over delayed shock conditions and of fixed over variable signal durations, it is assumed that preparatory responses must be precisely timed to coincide with shock (e.g., Knapp et al., 1959; Safarjan & D'Amato, 1977). Presumably, longer delays to shock make precise timing difficult. Yet to account for the stronger preference obtained with long over short signal durations, the opposite assumption is made, namely, that longer signals allow more effective preparation. Obviously, the precise timing and better preparation assumptions make opposite predictions in the same situation. The only criterion for choosing one assumption over the other appears to be the specific experimental outcome. A similar conflict of assumptions occurs when the rationale for choice is examined for studies involving the dependability of shock given the presence of a signal, or the dependability of no shock given the absence of a signal (Badia et al., 1976). In this case, it must be assumed that preparation is either effortful or effortless, or that there either is or is not a relation between preference and the pro-

portion of shocks to which preparation is made. In brief, when an internally consistent set of assumptions is adopted, coverage of the data is substantially restricted for the preparation hypothesis.

In addition to logical difficulties, there are also empirical difficulties for the preparation hypothesis (Badia, Coker, & Harsh, 1973; Badia, Culbertson, & Harsh, 1973). As noted earlier, subjects in the experiments of Badia and his colleagues chose signaled over unsignaled shock even though signaled shock was two to nine times longer, two to three times more intense, or four to eight times more dense. It seems unlikely that preparation would have reduced the aversiveness of signaled shock to that of unsignaled shock under all of these conditions. It is difficult to imagine a preparatory response so effective as to lower the aversiveness of the longer, stronger, or more dense signaled shock below that of the shorter, weaker, or less dense unsignaled shock. The results of the Harsh and Badia (1975) study showing that preference for the signaled condition increases as shock intensity increases are also difficult to reconcile with the preparation hypothesis. Presumably, preparation should occur at all intensity values. Other data argue against the preparation hypothesis as it relates to specific skeletal responses. The results of Miller et al. (1974), of Fisher and Badia (1975), and of others using surface electrodes to deliver shock rule out skeletal preparation that would have resulted in the subjects receiving different amounts of shock. Similarly, the study by Badia and Abbott (in press) monitoring shock duration found no differences between signaled and unsignaled conditions.

Some investigators have measured the response to an aversive event presented alone or preceded by a signal. Furedy and Doob (1972) summarized a number of experiments involving 150 human subjects and concluded that signaled shock failed to be rated less aversive than unsignaled shock. Similar results were also reported by Furedy and Ginsburg (1973) and Furedy and Klajner (1972).

An important series of studies reported by Gormezano and Coleman (1973) also relates directly to the question of preparation. Contrary to the preparation view, their findings suggest that the principle of reinforcement does not apply to classical conditioning. Gormezano and Coleman varied the effectiveness of preparation by attenuating the intensity of the US on trials in which a conditioned response was elicited. According to the preparation view, the frequency of conditioned responses should have increased; instead, this manipulation resulted either in no change or in a reduction in the frequency of conditioned responses.

Although a substantial literature indicates that a signal preceding shock attenuates an animal's distress vocalizations to that shock (e.g., Badia, Culbertson, Defran, & Lewis, 1971), the attenuation occurs the first time that the signal and shock are paired. Finding differences such as these on the first trial suggests that nonassociative factors are involved. Therefore the findings of Badia et al. cannot be used to support the preparation hypothesis. There are also studies with human subjects showing that the galvanic skin response to shock is smaller when shock is signaled (e.g., Baxter, 1966; Kimmel, 1967). This diminution in the galvanic skin response has also been found in the rat and has been interpreted as "preception" by Lykken (1962). Again, however, Badia and Defran (1970) have shown that the larger galvanic skin responses occurring to unsignaled shock resulted from orienting responses (Sokolov, 1963) occurring to the omission of the signal.

Safety Hypothesis

The first version of the safety hypothesis was offered by Mowrer (1960) to account for rats choosing signaled over unsignaled avoidable or escapable shock. Subsequently, the hypothesis was also used by Lockard (1963) and Seligman (1968). Seligman, Maier, and Solomon (1971) were the first to describe the analysis as the safety hypothesis, and they were primarily responsible for its development. Badia, Culbertson, and Lewis (1971) began to systematically apply the hypothesis to a wide range of preference findings, and the safety hypothesis soon challenged the preparation hypothesis as an interpretive model of preference in aversive situations.

The safety hypothesis emphasizes that situations with aversive stimuli can be divided into discriminably different components that vary in their degree of aversiveness. In its most simple form, emphasis is placed on discriminable shock and shock-free periods, whether shock stimulation is avoidable, escapable, or inescapable. These discriminable periods are orthogonal. For example, when subjects are given a choice between signaled and unsignaled shock, three distinct stimulus conditions exist: (a) the presence of the signal (CS) in the signaled condition, (b) the absence of the signal ($\overline{\mathrm{CS}}$) in the signaled condition, and (c) the unsignaled condition. Although the same shock distribution is used for both shock conditions and overall shock rates are identical, local shock rates may vary. In the signaled condition, shock always occurs in the presence of the signal, $p\ (\mathrm{US}|\mathrm{CS}) = 1.0$, and never in the absence of the signal, $p\ (\mathrm{US}|\overline{\mathrm{CS}}) = 0$. Under the signaled condition, therefore, shock (unsafe) periods and shock-free (safe) periods are perfectly identified, even when shock is randomly programmed. In contrast to the signaled condition, neither safe nor unsafe periods can be identified in the unsignaled condition, and with random shock, the entire intershock interval may acquire properties of an unsafe period. Thus, when subjects are in the signaled condition, the shock period is identifiable and usually brief—at most lasting only as long as the signal duration. Further, the safe period is also identifiable and considerably longer, since it consists of the total intershock times minus the signal duration. Presumably, subjects choose the signaled shock condition over the unsignaled one on this basis; that is, the safe periods are identifiable and are considerably longer than the unsafe periods.

Advantages of the safety hypothesis. One advantage of the safety analysis is that it can reconcile such findings as the acquired aversiveness of preshock stimuli with the findings that subjects prefer situations that include these stimuli. The safety hypothesis permits the generalization that stimuli paired with a reinforcer, positive or negative, acquire the properties of that reinforcer. Findings dealing with conditioned fear (e.g., McAllister & McAllister, 1971), with conditioned suppression (e.g., see review by Davis, 1968), and with inhibition and facilitation of avoidance (e.g., Rescorla & LoLordo, 1965) are compatible with this view. Support for the safety analysis can also be inferred from studies showing that greater physiological deterioration usually occurs when shock is unpredictable. Other data also support the safety analysis.

If safety is important, then its duration should be a factor. Stimuli associated with increased durations of shock-free time should acquire differential control over responses that produce these durations. The latter point is important because it demonstrates that factors other than the parameters of the aversive stimulus are controlling behavior. An example of one such effect involves the intertrial interval in avoidance learning, a factor known to be important in avoidance learning (e.g., Weisman & Litner, 1969a, 1969b). A more compelling set of data has also been provided by these investigators (Weisman & Litner, 1971). They used a procedure introduced by Rescorla and LoLordo (1965) in which a stimulus paired with shock or a different stimulus paired with no shock was presented. They systematically varied the duration of the no-shock period across groups and then imposed the stimuli identifying shock and no-shock periods on an avoidance task. When the stimulus identifying safety was imposed on avoidance, responding decreased. More important, the greater the duration of safety during the initial training, the greater the decrement in avoidance responding during testing. Other studies have also shown the duration of the shock-free period to be important (e.g., Azrin, Hake, Holz, & Hutchinson, 1965). As described earlier, Harsh and Badia (1976) demonstrated that the longer the duration of the shock-free period, the stronger the preference for the signaled over the unsignaled condition.

The results of Badia, Culbertson, and Lewis (1971) and of Badia and Culbertson (1972) also support a safety analysis. These investigators analyzed the stimuli within a preference task that controlled choosing the signaled condition. They demonstrated through a series

of extinction trials that the stimulus that was correlated with the shock-free period controlled changing from the unsignaled to the signaled condition. They also demonstrated that the stimulus (signal) that was correlated with the shock period, to which preparation could be made, did not maintain changing to the signaled condition. Other data also support the safety analysis. Harsh and Badia (1975) found that preference for the signaled condition varied with the intensity of shock. In another study (Harsh & Badia, 1974), they found that although subjects preferred the signaled shock condition, responding for food was most suppressed in the presence of the signal and least suppressed in its absence.

Difficulties with the safety hypothesis. Several findings are difficult for the safety analysis to accommodate. One of these findings deals with signal duration. Abbott and Badia (in press) found that animals did not prefer the signaled condition with signal durations of .5 or 1.0 sec but that with longer durations they did. Signal durations of the shorter length are clearly discriminable, thus allowing shock and shock-free periods to be identified; yet animals did not change to the signaled condition. Also, the finding that longer rather than shorter signal durations resulted in stronger preference for the signaled condition is incompatible with safety (French et al., 1972; Perkins et al., 1966). The recent work of D'Amato and Safarjan (1979) is also relevant. These investigators found that rats preferred information about the duration of shock that they were to receive, even though this information was unrelated to shock and shock-free periods. However, Freeman and Badia (1975), in a study similar to D'Amato and Safarjan, found that information about shock intensity did not result in a preference. Other results indicate that safety is not a necessary condition for preference. In a study by Badia et al. (1976), the dependability of the stimuli identifying shock and shock-free periods was varied. Preference for the signaled condition was maintained even when a number of unpredictable shocks were delivered during the formerly shock-free (safe) period. Only when the dependability of safety was reduced to a relatively low level was the preference eliminated. It is difficult for the safety analysis to deal with these data because the shock-free period was only relatively safe, not totally safe, under this latter condition.

Clearly, evidence for and against both the preparation analysis and the safety analysis exists, and perhaps both views have their merits under specific experimental conditions. However, preference for signaled shock may be determined by a number of factors, and it may not be possible to incorporate all relevant data under a single principle.

Prospectus

Our review of the literature has firmly established the reliability of preference for predictable shock situations under a variety of situations. The review rules out explanations of the phenomenon based on methodological considerations, such as differential shock avoidance or shock attenuation. It also rules out theoretical explanations of preference based on conditioned reinforcement or on the inherent value of information. Evidence favoring the remaining theoretical views of preparation and safety has been noted. We have also noted evidence incompatible with each of these latter views. For preparation, assumptions about the properties of preparatory responses across various conditions often are in conflict when viewed together. In addition, preparation theory fails to adequately account for the evidence showing that signals acquire aversive properties. The preparation view also has difficulty accounting for the preference for stronger, longer, or more dense signaled shock, for the controlling influence of stimuli correlated with safe periods, and for the failure of explicitly reinforced preparatory responses to maintain or strengthen conditioned responses. Similarly, the safety hypothesis has difficulty accounting for preference when safety is not completely dependable and for the marked influence that signal duration has on preference. It is apparent that neither the preparation hypothesis nor the safety hypothesis alone is sufficient to account for the available data. Predictable and unpredictable shock situations are probably more complex than these relatively simple theories

suggest. There may be other factors within these situations that are important.

One factor that may be important is local reduction in shock frequency or probability. Gibbon (1972, 1977), Herrnstein and Hineline (1966), Hineline (1970), and others have shown that reduction in shock density, both overall and local, is sufficient to maintain operant responding. Reduction in shock density also maintains choice behavior (e.g., Badia, Coker, & Harsh, 1973). Our view of preference based on shock-density reduction suggests that subjects change to a signaled condition because doing so frequently transfers them from a relatively high-density shock condition (unsignaled condition) to a relatively low-density shock condition (signal-absent component of signaled schedule), even though overall shock density is the same for both conditions. This view is compatible with a safety analysis based on discriminable periods relatively free of shock. When stated in relative rather than absolute terms, the safety hypothesis can account for choice performance under different dependabilities of safety. It is obvious that relative safety and local reduction in shock density are simply different ways of referring to the same controlling variable.

A second factor that may be important is the role played by classical conditioning. Even though this factor is explicitly recognized by the preparation hypothesis, it is treated only as a manifestation of the law of effect. Classical conditioning may have implications for behavior and physiology that differ from those stated by the preparation hypothesis. For example, whereas the preparation view suggests that conditioning should have beneficial effects due to a reduction in the aversiveness of the shock, the physiological evidence suggests that predictability can be more debilitating under certain conditions and less so under other conditions (Brady et al., 1962). More attention needs to be given to the role of classical conditioning and its interaction with operant choice behavior.

A third factor that needs attention is contrast. The signaled schedule may be compared to a multiple schedule in which the two components are identified by the signal's presence or absence. When components offering different reinforcement parameters are present in alternation, contrast effects often become evident. Indeed, the reinforcing value of safe periods emerges only as such periods are contrasted with periods of danger. The literature on contrast is extensive, but it is not yet clear how contrast alters the value and the physiological effects of signaled schedules.

Whatever the eventual theory, it is evident that signaled shock schedules are more complex than previously thought. Classical conditioning, successive contrast effects, local reinforcement, in addition to other factors, may all affect preference in signaled shock situations. The solution to the puzzle is not yet at hand, but progress toward such a solution is clearly being made.

Reference Notes

1. Furedy, J. J., & Walters, G. C. *Preference for signaled, supposedly unmodifiable shock as a function of scrambling the grid.* Paper presented at the meeting of the Psychonomic Society, San Antonio, Texas, November 1970.
2. Abbott, B., & Badia, P. *Choosing signaled shock: Some answers to recent criticisms.* Paper presented at the meeting of the Psychonomic Society, Washington, D.C., November 1977.

References

Abbott, B., & Badia, P. Choice for signaled over unsignaled shock as a function of signal length. *Journal of the Experimental Analysis of Behavior,* in press.

Arabian, J. M., & Desiderato, O. Preference for signaled shock: A test of two hypotheses. *Animal Learning & Behavior,* 1975, *3,* 191–195.

Ayers, J., Benedict, J. O., Glackenmeyer, R., & Matthews, W. Some factors involved in the comparison of response systems: Acquisition, extinction, and transfer of head-poke and lever-press Sidman avoidance. *Journal of the Experimental Analysis of Behavior,* 1974, *22,* 371–379.

Azrin N. H., Hake, D. V., Holz, W. C., & Hutchinson, R. R. Motivational aspects of escape from punishment. *Journal of the Experimental Analysis of Behavior,* 1965, *8,* 31–44.

Badia, P., & Abbott, B. Does shock modifiability contribute to preference for signaled shock? *Animal Learning & Behavior,* in press.

Badia, P., Coker, C. C., & Harsh, J. Choice of higher density signaled shock over lower density unsignaled shock. *Journal of the Experimental Analysis of Behavior,* 1973, *20,* 47–55.

Badia, P., & Culbertson, S. Behavioral effects of

signaled versus unsignaled shock during escape training in the rat. *Journal of Comparative and Physiological Psychology,* 1970, *72,* 216–222.

Badia, P., & Culbertson, S. The relative aversiveness of signaled vs. unsignaled escapable and inescapable shock. *Journal of the Experimental Analysis of Behavior,* 1972, *17,* 463–471.

Badia, P., Culbertson, S. A., Defran, R. H., & Lewis, P. Attenuation of rat vocalizations to shock by a stimulus: Sensory interaction effects. *Journal of Comparative and Physiological Psychology,* 1971, *76,* 131–136.

Badia, P., Culbertson, S., & Harsh, J. Choice of longer or stronger signaled shock over shorter or weaker unsignaled shock. *Journal of the Experimental Analysis of Behavior,* 1973, *19,* 25–32.

Badia, P., Culbertson, S. A., & Lewis, P. The relative aversiveness of signaled vs. unsignaled avoidance. *Journal of the Experimental Analysis of Behavior,* 1971, *16,* 113–121.

Badia, P., & Defran, R. H. Orienting responses and GSR conditioning: A dilemma. *Psychological Review,* 1970, *77,* 171–181.

Badia, P., & Harsh, J. Further comments concerning preference for signaled shock conditions. *Bulletin of the Psychonomic Society,* 1977, *10,* 17–20. (a)

Badia, P., & Harsh, J. Preference for signaled and unsignaled shock schedules: A reply to Furedy and Biederman. *Bulletin of the Psychonomic Society,* 1977, *10,* 13–16. (b)

Badia, P., Harsh, J., & Coker, C. C. Choosing between fixed time and variable time shock. *Learning and Motivation,* 1975, *6,* 264–278.

Badia, P., Harsh, J., Coker, C. C., & Abbott, B. Choice and the dependability of stimuli that predict shock and safety. *Journal of the Experimental Analysis of Behavior,* 1976, *26,* 95–111.

Bassett, J. R., Cairncross, K. D., & King, M. G. Parameters of novelty, shock predictability, and response contingency in corticosterone release in the rat. *Physiology and Behavior,* 1973, *10,* 901–907.

Baxter, R. Diminution and recovery of the UCR in delayed and trace classical GSR conditioning. *Journal of Experimental Psychology,* 1966, *71,* 447–451.

Berlyne, D. E. *Conflict, arousal, and curiosity.* New York: McGraw-Hill, 1960.

Biederman, G. B., & Furedy, J. J. The preference-for-signaled shock phenomenon: Effects of shock modifiability and light reinforcement. *Journal of Experimental Psychology,* 1973, *100,* 380–386.

Biederman, G. B., & Furedy, J. J. Operational duplication without behavioral replication of changeover for signaled inescapable shock. *Bulletin of the Psychonomic Society,* 1976, *7,* 421–424. (a)

Biederman, G. B., & Furedy, J. J. Preference for signaled shock in rats? Instrumentation and methodological errors in the archival literature. *Psychological Record,* 1976, *26,* 501–514. (b)

Biederman, G. B., & Furedy, J. J. The preference-for-signaled shock phenomenon: Fifty days with scrambled shock in the shuttlebox. *Bulletin of the Psychonomic Society,* 1976, *7,* 129–132. (c)

Blackman, D. Conditioned suppression and the effects of classical conditioning on operant behavior. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Brady, J. P., Thornton, D. R., DeFisher, D. Deleterious effects of anxiety elicited by conditioned pre-aversive stimuli in the rat. *Psychosomatic Medicine,* 1962, *24,* 590–595.

Brimer, C. J., & Kamin, L. J. Fear of the CS in avoidance training and fear from a sense of helplessness. *Canadian Journal of Psychology,* 1963, *17,* 188–193.

Church, R. M. Response suppression. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior.* New York: Appleton-Century-Crofts, 1969.

Collier, A. C. Preference for shock signals as a function of the temporal accuracy of the signals. *Learning and Motivation,* 1977, *8,* 159–170.

Coppock, H. W. Stimuli preceding electric shock can acquire positive reinforcing properties. *Journal of Comparative and Physiological Psychology,* 1954, *47,* 109–113.

Cotsonas, P. M. *Preference for signaled over unsignaled aversive stimulation.* Unpublished master's thesis, University of North Carolina at Chapel Hill, 1972.

Crabtree, M. S., & Kruger, B. M. Free choice of scrambled electric shock with rats. *Bulletin of the Psychonomic Society,* 1975, *6,* 352–354.

D'Amato, M. R., & Safarjan, W. R. Preference for information about shock duration. *Animal Learning & Behavior,* 1979, 7, 89–94.

Davis, H. Conditioned suppression: A survey of the literature. *Psychonomic Monograph Supplements,* 1968, *2*(14, Whole No. 30).

Davis, H., & McIntire, R. W. Conditioned suppression under positive, negative, and no contingency between conditioned and unconditioned stimuli. *Journal of the Experimental Analysis of Behavior,* 1969, *12,* 633–640.

Davis, H., Memmott, J., & Hurwitz, H. M. B. Effects of signals preceding and following shock on baseline responding during a conditioned-suppression procedure. *Journal of the Experimental Analysis of Behavior,* 1976, *25,* 263–277.

Defran, R. H. Reinforcing effects of stimuli paired with schedules of aversive control (Doctoral dissertation, Bowling Green State University, 1972). *Dissertation Abstracts International,* 1972, *33,* 1865B–2419B. (University Microfilms No. 72-27,218)

Denny, M. R. Relaxation theory and experiments. In F. R. Brush (Ed.), *Aversive conditioning and learning.* New York: Academic Press, 1971.

Dinsmoor, J. S., Browne, M. P., & Lawrence, C. E. A test of the negative discriminative stimulus as a reinforcer of observing. *Journal of the Experimental Analysis of Behavior,* 1972, *18,* 79–85.

Dinsmoor, J. A., Flint, G. A., Smith, R. G., & Viemeister, N. F. Differential reinforcing effects of stimuli associated with the presence or absence of

a schedule of punishment. In D. P. Hendry (Ed.), *Conditioned reinforcement*. Homewood, Ill.: Dorsey Press, 1969.

Fisher, C., & Badia, P. Preference for signaled or unsignaled shock in goldfish. *Bulletin of the Psychonomic Society*, 1975, *6*, 195–197.

Frankel, P. W., & Vom Saal, W. Preference for predicted over unpredicted shock. *Quarterly Journal of Experimental Psychology*, 1976, *28*, 441–447.

Freeman, J., & Badia, P. Do rats prefer information about shock intensity? *Bulletin of the Psychonomic Society*, 1975, *6*, 75–78.

French, D., Palestino, D., & Leeb, C. Preference for a warning in an unavoidable shock situation: Replication and extension. *Psychological Reports*, 1972, *30*, 72–74.

Friedman, S. B., & Ader, R. Parameters relevant to the experimental production of "stress" in the mouse. *Psychosomatic Medicine*, 1965, *27*, 27–30.

Furedy, J. J., & Biederman, G. B. Preference-for-signaled-shock phenomenon: Direct and indirect evidence for modifiability factors in the shuttle box. *Animal Learning & Behavior*, 1976, *4*, 1–5.

Furedy, J. J., & Doob, A. N. Signaling unmodifiable shocks. Limits on human informational cognitive control. *Journal of Personality and Social Psychology*, 1972, *21*, 111–115.

Furedy, J. J., & Ginsburg, S. Effects of varying signaling and intensity of shock on an unconfounded and novel electrodermal autonomic index in a variable and long interval classical conditioning paradigm. *Psychophysiology*, 1973, *4*, 328–334.

Furedy, J. J., & Klajner, F. Unconfounded autonomic indexes of the aversiveness of signaled and unsignaled shocks. *Journal of Experimental Psychology*, 1972, *92*, 313–318.

Gibbon, J. Timing and discrimination of shock density in avoidance. *Psychological Review*, 1972, *79*, 68–92.

Gibbon, J. Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, 1977, *24*, 279–325.

Gliner, J. A. Predictable vs. unpredictable shock: Preference behavior and stomach ulceration. *Physiology and Behavior*, 1972, *9*, 693–698.

Gormezano, I., & Coleman, S. R. The law of effect and CR contingent modification of the UCS. *Conditional Reflex*, 1973, *8*, 41–56.

Griffin, P., Honaker, L. M., Jones, D. E., & Pynes, L. T. Preference for signaled vs. unsignaled shock in pigeons with implanted electrodes. *Bulletin of the Psychonomic Society*, 1974, *4*, 141–143.

Harsh, J. Preference for signaled shock: A well established and reliable phenomenon. *Psychological Record*, 1978, *28*, 281–289.

Harsh, J., & Badia, P. A concurrent assessment of the positive and negative properties of a signaled shock schedule. *Animal Learning & Behavior*, 1974, *2*, 169–172.

Harsh, J., & Badia, P. Choice for signaled over unsignaled shock as a function of shock intensity. *Journal of the Experimental Analysis of Behavior*, 1975, *23*, 349–355.

Harsh, J., & Badia, P. A temporal parameter influencing choice between signaled and unsignaled shock schedules. *Journal of the Experimental Analysis of Behavior*, 1976, *25*, 327–333.

Herrnstein, R. J. Aperiodicity as a factor in choice. *Journal of the Experimental Analysis of Behavior*, 1964, *7*, 179–182.

Herrnstein, R. J., & Hineline, P. N. Negative reinforcement as shock frequency reduction. *Journal of the Experimental Analysis of Behavior*, 1966, *9*, 421–430.

Hershiser, D., & Trapold, M. A. Preference for unsignaled over signaled direct reinforcement in the rat. *Journal of Comparative and Physiological Psychology*, 1971, *77*, 323–328.

Hineline, P. N. Negative reinforcement without shock reduction. *Journal of the Experimental Analysis of Behavior*, 1970, *14*, 259–268.

Holmes, P. A., Jackson, D. E., & Byrum, R. P. Acquisition and extinction of conditioned suppression under two training procedures. *Learning and Motivation*, 1971, *2*, 334–340.

Hyman, A. Two temporal parameters of free operant discriminated avoidance in the rhesus monkey. *Journal of the Experimental Analysis of Behavior*, 1969, *12*, 641–648.

Hymowitz, N. Preference for signaled electric shock. *Proceedings of the 81st Annual Convention of the American Psychological Association*, 1973, *8*, 847–848. (Summary)

Hymowitz, N. Suppression of responding during signaled and unsignaled shock. *Psychological Bulletin*, 1979, *86*, 175–190.

Jenkins, H. M., & Boakes, R. A. Observing stimulus sources that signal food or no food. *Journal of the Experimental Analysis of Behavior*, 1973, *20*, 197–207.

Keehn, J. D. The effect of a warning signal on unrestricted avoidance behaviour. *British Journal of Psychology*, 1959, *50*, 125–135.

Kendall, S. B. Effects of two procedures for varying information transmission on observing responses. *Journal of the Experimental Analysis of Behavior*, 1973, *20*, 73–83.

Kimmel, E. Judgments of UCS intensity and diminution of the UCR in classical GSR conditioning. *Journal of Experimental Psychology*, 1967, *73*, 532–543.

Knapp, R. D., Kause, R. H., & Perkins, C. C., Jr. Immediate versus delayed shock in T-maze performance. *Journal of Experimental Psychology*, 1959, *58*, 357–362.

Lewis, P., & Gardner, E. T. The reliability of preference for signaled shock. *Bulletin of the Psychonomic Society*, 1977, *9*, 135–138.

Lockard, J. S. Choice of warning signal or no warning signal in an unavoidable shock situation. *Journal of Comparative and Physiological Psychology*, 1963, *56*, 526–530.

Lockard, J. S. Choice of a warning signal or none in several unavoidable-shock situations. *Psychonomic Science*, 1965, *3*, 5–6.

Logan, F. A., & Boice, R. Avoidance of a warning signal. *Psychonomic Science,* 1968, *13,* 53–54.

Lutz, R. E., & Perkins, C. C. A time variable in the acquisition of observing responses. *Journal of Comparative and Physiological Psychology,* 1960, *53,* 180–182.

Lykken, D. T. Preception in the rat: Autonomic response to shock as function of length of warning interval. *Science,* 1962, *137,* 136–137.

Marlin, N. A., Berk, A. M., & Miller, R. R. Modification and avoidance of unmodifiable and unavoidable footshock. *Bulletin of the Psychonomic Society,* 1978, *11,* 203–205.

McAllister, W. R., & McAllister, D. E. Behavioral measurement of conditioned fear. In F. R. Brush (Ed.), *Aversive conditioning and learning.* New York: Academic Press, 1971.

Mezinskis, J., Gliner, J., & Shemberg, K. Somatic response as a function of no signal, random signal, or signaled shock with variable or constant durations of shock. *Psychonomic Science,* 1971, *25,* 271–272.

Miller, R. R., Daniel, D., & Berk, A. M. Successive reversals of a discriminated preference for signaled tailshock. *Animal Learning & Behavior,* 1974, *2,* 271–274.

Miller, R. R., Marlin, N. A., & Berk, A. M. Reliability and sources of control of preference for signaled shock. *Animal Learning & Behavior,* 1977, *5,* 303–308.

Mitchell, K. M., Perkins, N. P., & Perkins, C. C., Jr. Conditions affecting acquisition of observing responses in the absence of differential reward. *Journal of Comparative and Physiological Psychology,* 1965, *60,* 435–437.

Mowrer, O. H. On the dual nature of learning—A re-interpretation of "conditioning" and "problem-solving." *Harvard Educational Review,* 1947, *17,* 102–148.

Mowrer, O. H. *Learning theory and the symbolic process.* New York: Wiley, 1960.

Nageishi, V., & Imada, H. Suppression of licking behavior in rats as a function of predictability of shock and probability of conditioned-stimulus-shock pairings. *Journal of Comparative and Physiological Psychology,* 1974, *87,* 1165–1173.

Paré, W. The effect of chronic environmental stress on stomach ulceration, adrenal function, and consumatory behavior in the rat. *Journal of Psychology,* 1964, *57,* 143–151.

Perkins, C. C., Jr. The stimulus conditions which follow learned responses. *Psychological Review,* 1955, *62,* 341–348.

Perkins, C. C., Jr. An analysis of the concept of reinforcement. *Psychological Review,* 1968, *75,* 155–172.

Perkins, C. C., Jr. Reinforcement in classical conditioning. In H. H. Kendler & J. T. Spence (Eds.), *Essays in neobehaviorism* New York: Appleton-Century-Crofts, 1971.

Perkins, C. C., Jr., Levis, D. J., & Seymann, R. Preference for signal-shock vs. shock-signal. *Psychological Reports,* 1963, *13,* 735–738.

Perkins, C. C., Jr., Seymann, R. G., Levis, D. J., & Spencer, H. R., Jr. Factors affecting preference for signal-shock over shock-signal. *Journal of Experimental Psychology,* 1966, *72,* 190–196.

Peterson, G. B., Ackil, J. E., Frommer, G. P., & Hearst, E. L. Conditioned approach and contact behavior toward signals for food or brain-stimulation reinforcement. *Science,* 1972, *177,* 1009–1011.

Powell, R. W. A comparison of signaled vs. unsignaled free-operant avoidance in wild and domesticated rats. *Animal Learning & Behavior,* 1976, *4,* 279–286.

Price, K. P. Predictable and unpredictable aversive events. Evidence for the safety signal hypothesis. *Psychonomic Science,* 1972, *26,* 215–216.

Prokasy, W. F. The acquisition of observing responses in the absence of differential external reinforcement. *Journal of Comparative and Physiological Psychology,* 1956, *49,* 131–134.

Rachlin, H., & Herrnstein, R. J. Hedonism revisited: On the negative law of effect. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior.* New York: Appleton-Century-Crofts, 1969.

Rescorla, R. A., & LoLordo, V. M. Inhibition of avoidance behavior. *Journal of Comparative and Physiological Psychology,* 1965, *59,* 406–410.

Safarjan, W. R., & D'Amato, M. R. Preference and information about the time and the occurrence of shock delivery. *Bulletin of the Psychonomic Society,* 1977, *10,* 355–357.

Safarjan, W. R., & D'Amato, M. R. Variables affecting preference for signaled shock in a symmetrical changeover design. *Learning and Motivation,* 1978, *9,* 314–331.

Schoenfeld, W. N. An experimental approach to anxiety, escape, and avoidance behavior. In P. H. Hoch & J. Zubin (Eds.), *Anxiety.* New York: Grune & Stratton, 1950.

Seligman, M. E. P. Chronic fear produced by unpredictable electric shock. *Journal of Comparative and Physiological Psychology,* 1968, *66,* 402–411.

Seligman, M. E. P., Maier, S. F., & Solomon, R. L. Unpredictable and uncontrollable aversive events. In F. R. Brush (Ed.), *Aversive conditioning and learning.* New York: Academic Press, 1971.

Seligman, M. E. P., & Meyer, B. Chronic fear and ulcers in rats as a function of the unpredictability of safety. *Journal of Comparative and Physiological Psychology,* 1970, *73,* 202–207.

Shimoff, E. H., Schoenfeld, W. N., & Snapper, A. A. G. Effects of CS presence and duration on suppression of positively reinforced responding in the rat. *Psychological Reports,* 1969, *25,* 111–114.

Sidman, M. Some properties of the warning stimulus in avoidance behavior. *Journal of Comparative and Physiological Psychology,* 1955, *48,* 444–450.

Simpson, C. W., Wilson, L. G. M., DiCara, L. V., Jarrett, K. J., & Carroll, B. J. Stress-induced ulceration in adrenalectomized and normal rats. *Bulletin of the Psychonomic Society,* 1975, *6,* 189–191.

Sokolov, E. N. *Perception and the conditioned reflex.* Oxford, England: Pergamon Press, 1963.
Ulrich, R. E., Holz, W., & Azrin, N. H. Stimulus control of avoidance behavior. *Journal of the Experimental Analysis of Behavior,* 1964, *7,* 129–133.
Weisman, R. G., & Litner, J. S. The course of Pavlovian excitation and inhibition of fear. *Journal of Comparative and Physiological Psychology,* 1969, *69,* 667–672. (a)
Weisman, R. G., & Litner, J. S. Positive conditioned reinforcement of Sidman avoidance behavior in rats. *Journal of Comparative and Physiological Psychology,* 1969, *68,* 597–603. (b)
Weisman, R. G., & Litner, J. S. Role of the intertrial interval in Pavlovian differential conditioning of fear in rats. *Journal of Comparative and Physiological Psychology,* 1971, *74,* 211–218.
Weiss, J. M. Somatic effects of predictable and unpredictable shock. *Psychosomatic Medicine,* 1970, *32,* 397–408.
Weiss, J. M. Effects of coping behavior in different warning signal conditions on stress pathology in rats. *Journal of Comparative and Physiological Psychology,* 1971, *77,* 1–13. (a)
Weiss, J. M. Effects of coping behavior with and without a feedback signal on stress pathology in rats. *Journal of Comparative and Physiological Psychology,* 1971, *77,* 22–30. (b)
Weiss, J. M. Effects of punishing the coping response (conflict) on stress pathology in rats. *Journal of Comparative and Physiological Psychology,* 1971, *77,* 14–21. (c)
Weiss, J. M. Ulcers. In J. D. Maser & M. E. P. Seligman (Eds.), *Psychopathology: Experimental models.* San Francisco: W. H. Freeman, 1977.
Weiss, K. M., & Strongman, K. T. Shock-induced response bursts and suppression. *Psychonomic Science,* 1969, *15,* 238–240.
Wilton, R. N., & Clements, R. O. Observing responses and informative stimuli. *Journal of the Experimental Analysis of Behavior,* 1971, *15,* 199–204.
Wyckoff, L. B., Jr. The role of observing responses in discrimination learning. Part I. *Psychological Review,* 1952, *59,* 431–442.

Received May 25, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1132–1148

The Alpha Experience Revisited: Biofeedback in the Transformation of Psychological State

William B. Plotkin
State University of New York at Albany

Presented is a review of empirical research and conceptual perspectives on the development of unusual experiential states during electroencephalographic (EEG) alpha-biofeedback training. It is concluded that the occurrence of the "alpha experience" is relatively independent of the strength or density of EEG alpha activity, and that the transformation in experience during feedback training can be accounted for by eight categories of complexly interrelated factors: (a) sensory deprivation, (b) sustained alertness, (c) concentration/meditation, (d) introspective sensitization, (e) expectation, (f) perceived success at the feedback task, (g) attribution processes, and (h) individual differences. Conceptual and empirical implications for biofeedback and for the study of physiological–experiential relationships are discussed.

During electroencephalographic (EEG) alpha-feedback training, trainees are presented with immediate moment-to-moment information on the strength or density of their EEG alpha rhythms, which provides them the opportunity, in principle, to learn to increase, maintain, or decrease the strength of this brain rhythm. Of special significance is the observation by some researchers (Brown, 1970; Hardt & Kamiya, 1976a; Hart, 1968; Kamiya, 1968, 1969; Nowlis & Kamiya, 1970) that many persons report entering a quasi-meditational state of consciousness during alpha-enhancement feedback training. This state of consciousness, often called the "alpha experience," is usually identified as a pleasant, relaxed, and serene state, characterized by a loss of body and time awareness, an absence or diminution of thought, and a feeling of egolessness (Brown, 1970; Hart, 1968; Kamiya, 1968, 1969; Nideffer, 1973; Nowlis & Kamiya, 1970; Plotkin, 1976a, 1977; Plotkin & Cohen, 1976; Walsh, 1974).

Initially, researchers claimed that the alpha experience was intrinsically and directly associated with enhanced alpha levels, and that the alpha experience was, in fact, *caused* by enhanced alpha levels (hence, the name, alpha experience). Frequently cited in support of this view (in addition to the early alpha-biofeedback studies) was the observation that the EEGs of meditators often show increased alpha strength during meditation (Anand, Chhina, & Singh, 1961; Kasamatsu & Hirai, 1969; Wallace, 1970). The possibility of directly influencing experience through voluntary control of the electrical activity of the brain was a rather provocative notion. Indeed, the great popular interest in biofeedback may have been primarily generated by the idea that brain wave–feedback training had the potential for being a more efficacious method than the traditional meditative disciplines for effecting meditative states, or that at least it was a method better suited to the "modern, Western temperament."

More recently, however, considerable doubt has arisen that changes in a person's EEG alpha level have any direct or simple relationship to the achievement of the alpha experience. Several studies have failed to find a significant occurrence of alpha experiences during alpha-enhancement feedback, especially when the research participants were not led to expect such an experience (Beatty,

Requests for reprints should be sent to William B. Plotkin, who is now at the Klamath Mental Health Center, 3314 Vandenberg Road, Klamath Falls, Oregon 97601.

1972; Lynch, Paskewitz, & Orne, 1974; Orne & Paskewitz, 1974; Peper, 1971; Plotkin, 1976a; Plotkin & Cohen, 1976; Regestein, Pegram, Cook, & Bradley, 1974; Travis, Kondo, & Knott, 1975). Plotkin (1976a), for example, found that one of his groups of research participants, who did not know which brain waves were being studied and who were not told what kind of experiences to expect, described experiences that showed "no consistent similarities with the experiences that have been widely associated with high and low alpha states" (p. 89). Travis et al. summarized the experiential reports of 140 persons who participated in four studies that examined the alpha-enhancement phenomenon. They concluded that the alpha-enhancement task is not as overwhelmingly pleasant as had been suggested by Nowlis and Kamiya (1970) and by Brown (1970).

On the other hand, there are several studies that have replicated the finding of alpha experiences during alpha training. However, when the relationship between EEG alpha and experience has been examined, these studies have uniformly failed to find significant correlations between the degree of alpha enhancement and the intensity or likelihood of alpha experiences (Beatty, 1972; Lynch et al., 1974; Plotkin, 1977; Plotkin, Mazer, & Loewy, 1976; Sacks, Fenwick, Marks, Fenton, & Hebden, 1972). Lynch et al., for example, concluded that their research participants' "largely positive reactions to the feedback procedure were not the result of large increases in alpha activity and are certainly not likely to have been a function of alpha activity levels alone" (p. 409). Using more formalized correlation techniques, Plotkin et al. found no correlation between the degree of alpha enhancement and the likelihood of an alpha experience.

Even more damaging to the thesis that the alpha experience is the result of alpha enhancement is the recently uncovered fact that there is absolutely no published evidence that alpha training has ever resulted in an unequivocal case of true alpha enhancement (Johnson, 1977; Paskewitz, 1977; Plotkin, 1978). That is, alpha levels have never been shown to rise above prefeedback eyes-closed resting baseline levels. The subbaseline increases in alpha production that have often been reported during alpha training have been shown to be the result of the gradual dissipation or neutralization of alpha-inhibitory influences, which is a case of disinhibition or habituation, not enhancement (Lynch & Paskewitz, 1971; Paskewitz, 1977; Paskewitz, Lynch, Orne, & Costello, 1970; Plotkin, 1978; Plotkin, Note 1).

Cognizant of this problem, Hardt and Kamiya (1976a) have argued that the failure to find reliable and significant alpha enhancement is due to the use of "deficient methodologies," such as insufficient training time or a percentage-of-time measure of alpha rather than an amplitude-integration measure (Hardt & Kamiya, 1976b). However, Plotkin (1976b) has pointed out that most of these "suspect" studies used methodologies that were similar to, or nearly identical with, those of the original studies of alpha-feedback training—those by Brown (1970), Kamiya (1968, 1969), and Nowlis and Kamiya (1970). Moreover, a recent study (Plotkin, 1978) that employed the precise methodology recommended by Hardt and Kamiya (1976a), including almost 9 hours of total training time, found no evidence for the learned enhancement of alpha strength significantly above optimal eyes-closed baseline levels, although in some cases alpha-enhancement training did result in the *maintenance* of optimal alpha levels.

In summary, there is now solid support for the conclusion that alpha-enhancement training per se is neither necessary for, nor especially facilitative of, the achievement of the alpha experience. However, it is important to note that the phenomenological authenticity of the alpha experience is not being called into question here. The point is that alpha enhancement per se has not been instrumental in—or intrinsic to—the achievement of this experience. Although there is always the problem of bias and compliance in the report of experiential states, most alpha researchers have learned that there is simply no doubt that many of their trainees have experienced highly unusual, meaningful, and occasionally profound alterations in consciousness during feedback training. That this is so is perhaps

Table 1
Factors Involved in the Development of Unusual Experimental States During Electroencephalograph Alpha-Biofeedback Training

1. Sensory deprivation due to
   (a) The biofeedback setting
   (b) Alpha-feedback-augmented sensory limitation
2. Sustained alertness (during sensory deprivation)
3. Concentration/meditation
4. Introspective sensitization
5. Suggestion and expectation due to
   (a) Preexperimental expectancies
   (b) Implicit suggestion
   (c) Explicit suggestion
6. Perceived success at the feedback task
7. Dual attribution of responsibility inherent in biofeedback training
8. Individual differences

best demonstrated by the extraordinary eagerness of many alpha trainees to repeat the experience, to learn all they can about it, and to spend considerable sums of money to purchase or rent the equipment that is seen as necessary for the generation of the experience (Lawrence, 1972). A recent study (Plotkin, in press) that employed a strong demand for honesty on experiential reports also supports this view. The question at this point is not whether these experiential reports are dismissable as artifacts, but rather, given that the attainment of the alpha experience during alpha training is not related to any unusual change in EEG alpha, how then *do* we explain the occurrence of these experiences?

We now appear to be in a position to offer an adequate answer to this question. Over the past few years there has accumulated a substantial body of evidence that demonstrates that there are at least eight categories of complexly interrelated factors (variables) that account for the occurrence of alpha experiences and similar states during alpha-biofeedback training. Table 1 presents an outline of these eight categories. Note that alpha enhancement is not among them.

Besides elucidating the development of alpha experiences, a review and discussion of these eight variables will highlight and illustrate several of the subtle, albeit critical, difficulties inherent in the attempt to establish direct or intrinsic relations between physiological states or processes and experiential or behavioral phenomena. There are complex methodological problems involved in determining how a particular physiological state is related, if at all, to a particular psychological state or behavior, and in deciding whether biofeedback or other means of altering particular physiological activities are critical to—or incidental to—the observed changes in psychological state (Shapiro, 1977). An understanding of how unusual experiential states are generated in the biofeedback setting will also enhance our knowledge of biofeedback training in its larger context of social and therapeutic influence—as opposed to its narrower definition as a method of facilitating physiological self-control.

Implicit in the reconstruction I shall offer of the development of these experiences will be a rejection of the reductionist and mechanist position that holds that psychological states are simply the consequence of efficient causes such as physiological processes and reinforcement histories. Instead, I shall proceed from a contextual human-action perspective, which recognizes that psychological state is one parameter or strand in a complex behavioral process that includes cognitions, motivations, and social significances as other parameters, as well as physiological processes and learning histories (Ossorio, 1973, 1978; Sarbin, 1977).

Sensory Deprivation

One hypothesis with respect to the development of unusual experiential states during alpha training is that the alpha-feedback setting happens to be conducive to the development of sensory deprivation and the associated alterations in consciousness (Zubeck, 1969). There appear to be, in fact, two independent aspects of alpha-training procedures that facilitate sensory deprivation: the attributes of the general biofeedback setting and the effects of alpha-enhancement training per se.

The Biofeedback Setting

There are several ways in which the typical alpha-biofeedback setting resembles those that

are employed in sensory-deprivation experimentation. Trainees are usually asked to sit in a comfortable chair or to lie on a bed, which is typically situated in a small soundproof or sound-attenuated room with low lighting or none at all. In addition, trainees are commonly asked to keep their eyes closed, to relax, and not to move around once they have become comfortable, in order not to disturb the EEG electrodes, which are sensitive to electromyographic (EMG) artifacts. Moreover, the standard feedback signal is a monotonous tone, usually appearing over a headphone set, which the trainees are constantly monitoring in order to track their changing alpha levels. Given these aspects of the typical alpha-feedback setting, it is not surprising that trainees often report becoming relaxed, with a loss of body awareness and with the associated sensory-deprivation feelings of lightness, floating, flying, or losing awareness of the "external" environment. However, as with sensory deprivation, some persons may react to this setting by falling asleep or with boredom, and some with anxiety or panic. The reason these latter responses are relatively rare during alpha training involves other factors discussed later, especially Factors 2 and 5 (see Table 1).

There has been only one piece of research (Plotkin, 1978) that has explicitly tested the hypothesis that the occurrence of the alpha experience is related to the sensory-deprivation aspects of the feedback setting. In this study, I found that persons who engaged in 10 52-min sessions of eyes-closed alpha-enhancement training without intrasession rest periods rated their experiences to be significantly more enjoyable and intense than did persons who engaged in precisely the same training with 20-sec eyes-open (and lights-on) rest periods interspersed every 4 min. The only significant difference in alpha levels between the two groups was a greater mean amplitude, on Session 1 only, for the group that did have the rest periods. Nevertheless, persons in the no-rest (high sensory-deprivation) group reported experiencing less body weight, greater personal involvement, faster speed of time, greater happiness, more emotional activation, greater personal relevance, more thought, and a "higher" state of consciousness. Moreover, relative to what might be thought to be more common procedures, the procedure of interspersing rest periods does not decrease the occurrence of alpha experiences: most of the studies that have reported the occurrence of these experiences have employed similar interspersed rest periods (Brown, 1970; Kamiya, 1968, 1969; Nowlis & Kamiya, 1970; Plotkin, 1976a, 1977; Walsh, 1974).

Alpha-Feedback-Augmented Sensory Limitation

Peper (1971) has noted another way in which the alpha experience may be related to a sensory-deprivation state. The research of Mulholland, his associates, and others (Mulholland, 1968, 1972, 1973; Mulholland & Peper, 1971; Wertheim, 1974) has demonstrated that the absence of occipital EEG alpha blocking reflects the absence of cortical oculomotor processing (in essence, abundant alpha occurs when a person is awake and "not looking"). Several other research reports (Chatrian, Magnus, Petersen, & Lazarte, 1959; Galin & Ornstein, 1972; Jasper & Penfield, 1949; Klass & Bickford, 1957; Kreitman & Shaw, 1965; Morgan, MacDonald, & Hilgard, 1974; Schwartz, Davidson, & Pugash, 1976) have suggested that the occurrence of alpha blocking at cortical locations other than the occipital lobe is also due to neural processing at the cortical location in question, with the concomitant activation of the behavioral processes associated with that location. In short, abundant alpha is known to accompany (or to be a sign of) sensory, motor, or cognitive quiescence at the cortical level. Thus, the *maintenance* of one's optimal *occipital* alpha level (which *is* facilitated through alpha-enhancement training; Plotkin, 1978; Note 1) would be expected to be accompanied by an absence of visual control processes. Moreover, because of the dominance of vision in the human being, alpha maintenance in the occipital lobe alone would be expected to have a *generalized* sensory limitation effect; that is, we would expect that the easiest way for an awake human being to minimize oculomotor activity—and thereby optimize

occipital alpha levels—would be for him to focus his attention on cognitive activity, and thus away from *all* sensory modalities, inasmuch as oculomotor activity is a concomitant of all sensory orientations.

In summary, it appears that occipital alpha-enhancement training results, to some degree, in a self-imposed sensory-deprivation state. Such a state would be expected to be characterized by increased nonsensory activity or awareness. However, the particular nature of this wakeful nonsensory state (e.g., introspection, daydreaming, boredom, hallucination, and some forms of meditation or contemplation) will depend on factors other than EEG alpha. Plotkin (1976a) and Plotkin and Cohen (1976) have demonstrated that there is a wide range of experiences that occur during alpha-enhancement training when the trainees are not led to expect any particular experiences. However, all of them are instances of nonsensory—and in particular, nonvisual—states.

Sustained Alertness During Sensory Deprivation

The quality of the experiences that occur during sensory deprivation would be expected to be very much influenced by the concurrent degree of alertness or drowsiness. In considering the explicit and implicit demands for relaxation in conjunction with the physical attributes of the feedback setting, one would expect drowsiness to regularly accompany alpha-feedback training. This development would be a problem for the researcher who is interested in evoking the alpha experience, since it cannot be experienced if the trainee is asleep; the alpha experience is an alert (although relaxed) state.

It is fortunate, therefore, that alpha-enhancement training facilitates the maintenance of alert wakefulness by facilitating the maintenance of naturally occurring eyes-closed alpha amplitudes. This feature of alpha training stems from one of the oldest EEG findings: Alpha activity decreases in amplitude and frequency, and essentially disappears, as a person becomes drowsy and approaches sleep (Adrian & Mathews, 1934; Berger, 1930; Lindsley, 1960). Thus, in order to keep the feedback signal on, the alpha trainee must learn to stay alert under sensory-deprivation conditions, a nontrivial task at which most trainees are nevertheless able to succeed. However, any task that would facilitate alertness and be compatible with sensory deprivation would do as well as alpha training in this regard. Yet we should note that alpha training is especially well suited for this purpose because it can be an absorbing task despite the fact that it involves only monotonous sensory stimulation (which renders it compatible with sensory deprivation).

Relaxed alertness may also be facilitated in the biofeedback setting by an upright posture (as in the traditional meditation position), by high levels of motivation or expectancy (to be discussed later), and, of course, by normal amounts of prior sleep.

The role of alpha training in facilitating relaxed alertness may help to explain why research participants in a noncontingent-feedback or a no-feedback group might not be as likely to report alpha experiences as those in a contingent-feedback group: The noncontingent and no-feedback participants are more likely to drowse off. Thus, it is not the case that enhanced alpha *causes* the alpha experience, or even that *maintained* optimal alpha is uniquely, intrinsically, or directly associated with the alpha experience; rather, drowsiness or sleep (which is accompanied by reduced alpha levels) is *in*compatible with the alpha experience. Maintained optimal alpha per se is as closely associated with alert daydreaming, mind-wandering, and boredom as it is with meditative experiences. Therefore, although alpha training per se may contribute to the occurrence of the alpha experience, it is not especially facilitative of it. The other factors discussed above and those to be examined below have been shown to be much more critical and influential in effecting the experiential state.

Concentration/Meditation

In addition to its sensory-deprivation qualities, the alpha-training procedure has some other important similarities to many meditation exercises, for example, immobility and concentration on, or sustained attention to, a

monotonous stimulus. The alpha trainees' task is to keep the alpha tone on as long (and/ or as loud) as possible. To accomplish this, they must intently focus their attention on the feedback tone, its variation, and the relation between the tone and their behavior and experience. This prolonged concentration on the feedback tone is formally equivalent to the meditator's sustained attention to breathing, to a mantra, to chanting or prayer, to a mandala, or to any other invariant or regular form (i.e., meditation object). As Naranjo and Ornstein (1971) have pointed out, this form of meditation exercise, which they call concentrative meditation, eventually results in a temporary suspension of ordinary thought, which is a central feature of the meditative (and alpha) experience, also reported by Deikman (1963) in his study of experimental meditation.

However, since the significance of the feedback signal, in its role as a meditation object, derives from its monotony, neutrality, and simplicity and not from its EEG contingency, it follows that a noncontingent tone would serve as well, in this regard, as the alpha tone in facilitating the generation of the alpha experience.

Introspective Sensitization

Recently, in a highly intriguing study, Hunt and Chefurka (1976) demonstrated that short periods of simply paying direct attention to one's "immediate subjective experience" elicited "anomalous subjective reports" and "altered-state effects" (p. 867). By "immediate subjective experience," Hunt and Chefurka mean "the bare features of momentary awareness without any reference to the consensual world of objects, persons, and meanings" (p. 868)—the "stimulus qualities" of sensations devoid of their significance as observations of everyday objects. Research participants who were requested simply to pay attention in this fashion for 10 min, without any explicit suggestions as to what to expect, generally reported "visual anomalies, uncanny emotion, . . . cognitive disorientation, and . . . feelings of interpersonal detachment and loneliness" (p. 872). The authors state that such data "suggest that altered-state effects can be tapped in very short time periods in any situation involving lack of movement, isolation, and at least implicitly, some attention to subjective experience" (p. 869).

The alpha-training situation certainly includes all of the latter three features. As we have seen, isolation and lack of movement are components of the sensory-deprivation qualities of the biofeedback setting. In addition, nearly all alpha-training studies include, at the very least, the implication that the training will result in mild to profound changes in experiential state. These experiential changes, which include changes in body awareness, are often explicitly outlined for the research participant before the onset of training (as will be discussed later). Thus we can conclude that the simple act of paying direct attention to one's sensations *as* sensations, which is a feature of the alpha-training context, can be expected to result to some degree in unusual experiential reports, independent of explicit suggestion, the degree of alertness, EEG alpha amplitudes, the presence of tones, or EEG-tone contingencies. "Introspective sensitization," as Hunt and Chefurka (1976) have termed it, is as much a feature of the alpha-training setting as it is of sensory deprivation, meditation, and hypnosis. Erickson, Rossi, and Rossi (1976), for example, have pointed out that

> the essential identity between periods of introspection and trance was demonstrated by Erickson . . . when he found that groups of subjects asked to perform a task in introspection underwent behavioral and subjective experiences that were similar to those they had when they went through a classical hypnotic induction. (p. 196)

Hunt and Chefurka (1976) also found that the experimental protocols of the classical introspectionists (e.g., Titchener, 1912; James, 1950; and Spearman, 1923) "revealed subjective anomalies similar to those found in drug and meditational states" (p. 867).

Suggestion and Expectation

The most widely endorsed hypotheses advanced to explain why unusual experiential states occur during alpha training have evoked such social psychological factors as suggestion, expectation, and the demand characteristics of the experimental setting (Beatty,

1972; DeGood, Elkin, Lessin, & Valle, 1977; Lynch & Paskewitz, 1971; Lynch et al., 1974; Peper, 1971; Plotkin, 1976a, 1976b, 1977, 1978; Plotkin & Cohen, 1976; Plotkin et al., 1976; Valle & Levine, 1975; Walsh, 1974; Glaros, Note 2). The central phenomenon here, of course, is the research participants' expectations about what sort of experiential changes will take place during training. These expectations can come about through (a) pre-experimental knowledge of alpha waves and/or alpha training, (b) explicit suggestion from the experimenter or from confederates, or (c) implicit suggestion (i.e., other demand characteristics).

It is certainly not surprising that expectation would play an important role in the development of alpha experiences during alpha-feedback training. After all, it has long been known that expectation has a very powerful influence on the often unusual experiences associated with hypnosis, relaxation procedures, meditation, and psychoactive drugs, as well as on all sorts of more common experiences. In reference to alpha training, Lynch and Paskewitz (1971), for instance, have made the following observation:

> Subjective reports are frequently influenced by the experimental setting and the course of the experiment itself. It is certainly possible that some of the reports of Ss in the feedback situation are influenced by what Orne (1962) has called the "demand characteristics" of the situation, that is, Ss enter the experiment expecting to experience alterations in mood, expecting the session to be pleasant, perhaps a "high," or if they don't feel this way initially, the experimenter may reinforce such feelings, both in the pre-experimental interview and in the actual instructions given during the experiment. (p. 212)

Preexperimental Expectancies

There has been only one published study (DeGood et al., 1977) that has explicitly assessed the effect of preexperimental expectancies on experiential reports of alpha training. DeGood et al. gave each of their research participants two 30-min feedback sessions: an alpha-enhancement session and an alpha-suppression session. Half of the participants had indicated on a screening questionnaire that they had some knowledge of alpha training, whereas the other half were "unknowledgeable" persons. The postexperimental experiential questionnaires indicated that only the knowledgeable persons experienced different subjective states during the enhancement and suppression sessions, with their reports of enhancement training being significantly more like the alpha experience than their reports of suppression training were.

Implicit Suggestion

To my knowledge, the effects of implicit suggestion in the alpha-feedback setting have never been documented independently of preexperimental expectancy. For instance, when expectancies are operative, one would naturally expect that informing a subject that she or he is about to begin an "enhancement trial" would serve as an implicit suggestion that the alpha experience is about to occur. This view offers the most plausible interpretation of a recent study by Glaros (Note 2), in which participants in one of the groups recruited for an "alpha-wave experiment" were given noncontingent (tape-recorded) feedback during both "enhancement" and "suppression" trials. These persons reported significantly more alpha experiences during the "enhancement" trials even though there was no difference in EEG alpha density between these two conditions.

Implicit suggestion should also be contrasted with inadvertent or informal suggestion, which occurs when the suggestion is not a formal component of the experimental instructions. Presumably, inadvertent or informal suggestion was at play in the early alpha studies, in which it appeared that there was a unique or intrinsic relation between alpha enhancement and the alpha experience (Brown, 1970; Kamiya, 1969; Nowlis & Kamiya, 1970).

Explicit Suggestion

The effects of explicit suggestion on reported experiences during alpha training have been an informal or secondary focus of several studies (e.g., Beatty, 1972; Lynch et al., 1974; Plotkin, 1976a). Beatty compared the experiential reports of participants in two contingent-feedback groups, one of which received no information about the alpha ex-

perience, while the other was informed of the phenomenological attributes of the experience. He reported a lack of uniformity in the reports of the no-information group. However, "subjects in the Information Condition, presumably because of their initial biases, reported the typical correlates of brain alpha rhythms—relaxation, calmness, inner awareness, etc." (p. 154). Lynch et al. reported that most of their subjects had positive reactions to the feedback procedures, but that "the most likely explanation for these positive reports rests in the fact that Ss were told that the experience would be a pleasant one" (p. 409).

There are three published studies that have systematically investigated the effects of explicit suggestion concerning the nature of the alpha experience (Plotkin, 1977; Plotkin et al., 1976; Walsh, 1974). Walsh employed a bidirectional design in which half of his research participants received alpha-suppression feedback and half received alpha-enhancement feedback. In addition, half of the persons in each of these groups received a "positive" alpha-experience set (explicit suggestion as to the nature of the experience), and half received a neutral set (general description of several possible experiences). Each research participant then received two 20-min sessions of alpha training, one with eyes open and one with eyes closed. Walsh found the typical alpha experience to be reported only when persons were given both the alpha-experience set and alpha-enhancement feedback. Either alone was not sufficient. Walsh interpreted these results as not only demonstrating the importance of suggestion, but also showing that the alpha experience is directly associated with the alpha rhythm, though this association may be blocked by "situational factors" unless the person is provided with "appropriate preparation for the experience, including some concepts to use in describing it" (p. 433). This latter conclusion, however, is not warranted by Walsh's data, since there is the following alternative interpretation, which has considerable independent support. Rather than demonstrating that the alpha experience is directly associated with alpha activity, Walsh's study may have shown only that the relative *absence* of alpha activity (during alpha-suppression feedback) is partially or wholly *incompatible* with the alpha experience (as well as with numerous other sorts of experiences associated with an inhibition of alpha blocking). According to this interpretation, the absence of alpha experiences in the groups that received alpha-suppression feedback (regardless of whether or not they also received the alpha-experience set) would be explained by this incompatibility rather than by a special, direct, or one-to-one relation between the alpha rhythm and the alpha experience. The finding of significantly more alpha experiences in the alpha-enhancement group that also received the alpha-experience set can be straightforwardly interpreted as demonstrating the effects of suggestion. Plotkin (1976a, 1977) and Plotkin and Cohen (1976) present data that demonstrate that alpha suppression, through its association with oculomotor activation, is antagonistic to the occurrence of the alpha experience.

In a related study, Plotkin et al. (1976) attempted to demonstrate the effects of suggestion on the experience of alpha training. Before the start of the 30-min eyes-open alpha-enhancement training, one group received an alpha-experience set; these research participants were explicitly informed about the specific experiential changes that were associated with increases in the volume of the feedback tone. The other group received no explicit suggestions whatever regarding what sort of experiences, if any, to expect. (Note that this no-set group differs from Walsh's neutral-set group. The latter was informed of a wide range of possible experiences, which included but did not emphasize the alpha experience.) No mention was made to any participant in either group that the research had anything to do with alpha waves (in order not to evoke preexperimental expectations). After the session, written experiential reports were collected and rated by blind judges on their similarity to a standardized description of the alpha experience. Somewhat surprisingly, the results showed that the likelihood of an alpha experience was not significantly different for the two

groups. However, this finding later came into focus when it was discovered that many of the research participants had spontaneously noted, on their postexperimental questionnaires, that they had felt very frustrated in their attempts to increase the tone volume (i.e., to enhance alpha). Thus, the alpha-experience set may have been ineffective at evoking alpha experiences because (a) the experience of frustration was incompatible with the alpha experience and (b) many trainees saw themselves as having failed at the very task that leads, they were told, to the alpha experience. This interpretation, then, suggests that the degree of *perceived* success at the feedback task should interact strongly with expectation. This hypothesis was tested in the following study.

Perceived Success at the Feedback Task

After the Plotkin et al. (1976) study, I was interested in demonstrating two separate points: (a) that the intensity of experiences reported to occur during alpha training would depend on perceived success at the enhancement task and be independent of actual success (the actual alpha amplitude relative to baseline) and (b) that the specific quality of the experiences would depend on the explicit suggestions given to the participants, assuming that preexperimental expectations were minimized. These two hypotheses were incorporated into a single 2 × 2 factorial design, in which the two between-group variables were Perceived Success (high or low) and Expectation (of one of two different sorts of altered states of consciousness). All participants received four 30-min. sessions of eyes-closed alpha-enhancement training, although it was insured that no participant thought that he or she was participating in research that was at all concerned with alpha waves. Persons who were randomly assigned to the lambda-expectation group were led to believe that they were training to enhance their "lambda" brainwaves, whereas the participants in the kappa-expectation group were told they were going to learn "kappa" enhancement. Both groups were told that they would find themselves in an altered state of consciousness (the "lambda" or "kappa" state, depending on the group) if they were successful at lambda (or kappa) enhancement. To determine the power of the Expectation variable, the lambda and kappa states were described as maximally dissimilar within the constraints of the biofeedback setting and the necessity of matching the motivational levels of the two groups. For instance, because of the sensory-deprivation qualities of the feedback setting, both groups were informed that the experience involved a loss of body awareness. Also, to insure an equal motivation to succeed, both states were described as pleasant, relaxing, and highly unusual and special. On the other hand, the two states were defined at opposite poles with respect to the more definitive aspects of the alpha experience: (a) the deliberateness and speed of thought, (b) the degree of "ego awareness," (c) emotionality, and (d) awareness of time. The lambda state (which corresponds to the alpha experience) was situated at the low end of each of these four dimensions, and the kappa state, at the high end. The major distinctive features of the lambda state were described as follows in the protocols:

> The "mind" slows down considerably during the lambda experience until the point is reached at which there is absolutely no thought, a condition often described as "blank mind." Even in the lighter stages . . . thought is very slow and free-flowing. . . . Eventually, thought stops entirely. However, the mind is nevertheless alert and awake at all times. . . . It is a clear and serene state—beyond emotion. . . . It is an "egoless" state, characterized by little or no awareness of oneself as a separate entity.

In contrast, the kappa state was described in the following manner:

> The "mind" becomes much more efficient in a certain sense. That is, the kappa experience is a state of very abundant and highly deliberate thought; we are able to direct our thought to any topic of interest and to process information at an extremely rapid rate. . . . It is characterized by abundant personal thought of a significant and often insightful nature. The kappa experience often involves thoughts of interpersonal experiences that are emotionally relevant . . . you become very aware of your personal strengths and of precisely who you are as a person.

Although all the research participants received the same form of contingent alpha feedback, half of those in each of the above groups—the "success" subjects—were made

to feel highly successful at the feedback task (regardless of the actual degree of success) by reporting to them (every 2 min via an intercom) a three-digit number that was ostensibly proportional to the preceeding trial's average alpha ("lambda" or "kappa") amplitude but was, in fact, the actual score inflated at a rate of an additional 2% every 2 min. Thus, while these "success scores" were still responsive to actual alpha amplitudes, and while the actual feedback tones were still being used, these trainees were nevertheless led to believe that they were improving at a somewhat remarkable, yet convincing, rate. In addition, persons in the Success group were given frequent verbal praise. On the other hand, persons in the "Failure" groups were given their actual 2-min scores, although their instructions informed them that the kappa (or lambda) experience does not even begin to occur until kappa (or lambda) strength is increased by at least 100% over initial levels (which never happens).

The fact that all participants were given contingent feedback is noteworthy. As many researchers have informally noted in their own labs, and as Strayer, Scott, and Bakan (1973) have formally demonstrated, many persons receiving noncontingent alpha feedback quickly grow discouraged, and often become drowsy or fall asleep. For this reason, noncontingent feedback (e.g., tape-recorded tones of a successful trainee) would not be effective for the Success group (they would not be as likely to feel successful) or for the Failure group (because, if they became drowsy, they would then manifest lower alpha amplitudes than the Success group, in which case perceived success would be confounded with actual success). In the experiment under consideration, then, the Failure subjects were able to feel that they had at least some control over the tone (which they did, in fact, have), although they were not able to produce the degree of enhancement that they believed was required to experience the alteration in consciousness. At any rate, it is because of the danger of the noncontingency being recognized that a noncontingent-feedback group is often not an adequate control group (Plotkin, Note 1).

The results from this study were straightforward. In general, and with few exceptions, in the written postexperimental questionnaires persons in the Kappa–Success group reported very powerful and genuine kappa experiences, while persons in the Lambda–Success group reported authentic lambda experiences. Most persons in both of the Failure groups reported "nothing unusual." These differences were borne out by statistical comparisons of the participants' ratings of their experiences on a series of 1-to-9 scales (Plotkin, 1977). As for the EEGs, there were no differences even approaching statistical significance between any of these groups in the degree of alpha enhancement actually achieved.

These results demonstrated (a) that the general intensity, pleasantness, value, and the degree of relaxation and sensory-deprivation effects that are reported to occur during alpha training are, indeed, very strongly influenced by the degree of perceived success at the feedback task; (b) that there is a wide range of experiences compatible with the biofeedback setting and abundant alpha activity—the "alpha experience" does not have a unique relationship to EEG alpha; (c) that the specific qualities of the reported experiences are closely related to the participants' expectations; and (d) that the quality and intensity of the experiences that occur during alpha training are independent of the actual degree of success at alpha enhancement.

The Attribution Process in Alpha Training

It has been established that alpha training per se is of little importance in the generation of experiential changes during these procedures. Nevertheless, it may be the case that the alpha experience is more likely to occur when the research participant can *attribute* the experience to alpha training or to another biofeedback procedure. For instance, would we find—or expect to find—equally profound changes in consciousness if we simply placed a person in a dark room for an hour and told him to expect—or to produce—certain experiential changes? What is it about the fact

that alpha trainees have this specific attribution available to explain to themselves the occurrence of these unusual experiences (namely, that they are the result of alpha training) that might enhance the likelihood of such experiences occurring in the first place?

An initial answer would be that alpha training serves essentially the same role as an inactive drug placebo: Just as there are many persons who will experience a suggested psychological effect after ingesting a purportedly psychoactive drug that is in fact only a sugar pill, there are many biofeedback trainees who will experience a suggested psychological effect during purportedly psychoactive biofeedback training that in fact has no significant physiological effect. (See Peek, 1977, for a discerning conceptualization of the placebo effect.)

However, the expectancy effect that is realizable in the biofeedback setting may be more powerful, or may at least have more active dimensions, than the typical drug placebo. There is one particular difference between a biofeedback treatment and a drug placebo that perhaps constitutes the most notable contribution of the entire biofeedback approach to therapeutic intervention: namely, the opportunity for the client, patient, or research participant to become an active agent in the process of change, control, or therapy (Stroebel & Glueck, 1973; Plotkin, Note 3). Whereas the recipients of a drug placebo are led to attribute the physiological, behavioral, or psychological transformation entirely to the drug (and thereby to reduce their own sense of responsibility and self-control), biofeedback trainees (whether or not the training per se has a humanly significant effect) will attribute a desirable outcome at least partially to themselves, which will *enhance* their sense of responsibility and self-control. In addition, there is, of course, a much greater likelihood of individuals eventually achieving complete self-control (without the aid of drugs or biofeedback) when they start from a point of *some* perceived control and move to one of more control, than when they attempt to go from no control to some control (Davison & Valins, 1969). In short, unlike placebo-treated persons, biofeedback trainees have been prepared to see themselves as least eligible for self-control of their problems, behavior, and/or experience (Plotkin, Note 3).

Thus, the biofeedback approach takes advantage of a combination of internal and external attributions of the suggested effect. There are two reasons why a biofeedback intervention may lead to a more powerful effect than an analogous external-placebo approach does. First is the fact that the experience of success at the feedback task may contribute directly to the outcome, especially when the major effect is an experiential or psychological change, as in the present case of alpha training. The self-control of an "involuntary" bodily process, especially one as mysterious and vital as brain wave activity, may be justifiable grounds for feelings of unusual self-mastery and for the accompanying positive affect. Furthermore, in the case of biofeedback there is nothing ambiguous about the occurrence of success: There is an objective measure of progress in the form of a feedback meter, tone, or other quantified index of control. Thus, the clinician or experimenter who employs biofeedback as a placebo intervention can arrange for his or her trainee to receive an indisputable feedback of "progress," which can serve as a very compelling counteragent to a trainee's lack of self-confidence and hence, as a powerful mobilizer of the trainee's motivations and skills.

The second advantage that the biofeedback intervention has over the external placebo follows from the fact that biofeedback trainees see themselves as active agents; they are therefore motivated to exert their own efforts toward producing the effect, an approach that may be expected to be more successful than that of placebo-treated persons, who usually have no reason to actively "help along" the drug (Valins & Nisbett, 1972). Alpha trainees would be expected to become more involved in—and thereby more influenced by—a procedure the effects of which they can see themselves as having facilitated than they would in a procedure that is ostensibly produced solely by an external agent.

Thus, Valins and Nisbett recommend that the individual who is treated with a drug or

placebo intervention be advised that the drug is "not so strong" and that it must be "helped along" by the appropriate self-control behaviors. When these procedures are used, the individual's self-doubts (about whether he or she can contribute to the production of the effect) are circumvented, and motivation and involvement are maintained. The biofeedback placebo goes even further in that the trainee's immediate task—control of the feedback signal—is at least one step removed from control of the target process or state (e.g., blood pressure, muscle group, or state of consciousness) and is thus less likely to evoke the trainee's doubts concerning his or her competence.

We would expect that alpha training would also be more effective than a procedure in which only *internal* attributions are available because most persons would probably not see themselves as able to induce such experiential states on their own without special training (for if they did, they would have done so already!). A person who starts out on a task that is believed to be impossible or doomed to failure is obviously less likely to succeed than one who thinks he or she has a good chance to succeed (Peek, 1977; Plotkin, Note 3).

These views concerning the attribution process in alpha training have received some support in a recently completed study (Plotkin, in press) in which experiential reports from six groups were compared. All the research participants were exposed to the identical physical setting and received the same explicit suggestions of an alpha experience. They were divided into the following 8 groups: (a) contingent EEG alpha-biofeedback training (participants in this group were instructed to try to increase the volume of a feedback tone; they were told that successful performance would enhance the strength of their alpha brain waves and thereby result in the alpha experience); (b) noncontingent biofeedback (instructions were identical to those above, but the "feedback" tone was in fact a tape recording of a successful trainee's feedback); (c) concentration exercise (an internal-attribution-only condition; the participants' task was to use the tape-recorded tone as a concentration object; successful concentration would lead to the alpha experience); (d) brain wave stimulation (the analog to the inactive-drug placebo; an external-attribution-only condition; the participants were informed that the alpha experience would be directly induced by a combination of electrical brain stimulation and computer-programmed auditory stimulation); (e) a combination of the last two groups (the participants were told that they were receiving direct brain wave stimulation, but that they must "help it along" by concentrating on the tone); and (f) self-induction (another internal-attribution-only condition, but unlike the concentration condition, participants were given no induction strategy; rather, they were on their own to self-induce the alpha experience in any way they could). The results from this study indicated that persons in the first two groups (biofeedback training) reported significantly more intense alpha experiences than did those in the other four groups. Moreover, there were no differences in experiential reports between the contingent and noncontingent versions of the biofeedback condition.

Individual Differences

Although the results of the Plotkin (1977) study demonstrated that reported experiences during alpha training were closely related to the participants' expectations and their degree of perceived success, there were nevertheless a few highly atypical responses in each group. A few persons in the Success groups reported no unusual experiences, or even the opposite experience to what they had been led to expect. In addition, a few persons in the Failure groups reported the suggested experiences despite our attempts to induce a perception of failure. These findings point to the importance of considering individual differences. Persons may have greater or lesser ability and/or disposition to self-induce unusual experiences or to self-induce one kind of experience over another. Moreover, persons differ in their proneness to experience sensory-deprivation effects, in their susceptibility or openness to suggestion, in their capacity to be comfortable in an experimental setting, in their disposition to follow instructions and

cooperate with the experimenter, and in many other relevant attributes. There has been very little work explicitly relating personal characteristics to the individual differences in reported experience during alpha training. A study in progress (Plotkin, Note 4) will correlate experimental reports with (a) state and trait anxiety (Spielberger, Gorsuch, & Lushene, 1970), (b) Rotter's (1966) Locus of Control Test, and (c) Shor's (1960) Personal Experience Questionnaire, which assesses the individual's proneness to naturally occurring altered-state experiences.

Related to the issue of individual differences is the study by Marshall and Bentler (1976), which demonstrated that alpha trainees who are deeply relaxed during training (as measured by forehead EMG) report significantly more alpha experiences than less relaxed trainees do, even when there are no differences in alpha density between the two groups.

Discussion

The research and concepts reviewed here appear to provide an adequate explanation of the development of unusual experiential states during alpha training. Of particular interest is the finding that there is no special, unique, or intrinsic relation between EEG alpha levels and the likelihood or intensity of the meditative state of consciousness known as the "alpha experience." Yet, at the same time, we can understand why it once appeared to researchers who were absorbed by the idea of a direct and simple relation between states of consciousness and neurophysiology that there was such an association: As it turns out, the alpha-feedback *situation* appears to be as effective as any other known procedure for generating such experiences in persons who do not have special meditation training. However, researchers were evidently unsuspecting of the strength and complexity of the eight non–EEG factors outlined in Table 1. Naturally, the early biofeedback investigators assumed that it was their operant-conditioning procedures, and not a conspiratorial set of "incidental" variables, that were responsible for their exciting results.

However, the research reviewed here does more than remind us that experiential states are complexly related to a host of parameters besides EEG alpha levels. It calls into question the entire enterprise of "mapping consciousness" neurophysiologically (Hilgard, 1969; Kamiya, 1968; Peper, 1972; Stoyva & Kamiya, 1968). As Grossberg (1972) has pointed out in reference to the present context of alpha-feedback studies, physiology and experience are of distinctly different logical types, so that not only is it unsurprising that we rarely find very tight relationships between them, it is also conceptually inappropriate to speak of any empirical correlations between them as a case of "mapping" if by this term we mean that the physiological states are formally equivalent to—or are efficient causes of—states of awareness, experiences, or behavior. What is called for at this point is not additional empirical specifications of physiology/experience or physiology/behavior correlations but rather an explicit and systematic articulation of the concepts of "persons" and "behavior," and of the logical relations between behavior, experience, psychological state, and physiological states (e.g., see Ossorio, 1973, 1978). Such an articulation will allow an understanding of the significance and implications of psychophysiological correlations that goes well beyond the simplistic and misleading notions of "mapping" or efficient cause.

There are several other implications of the findings reviewed here. First, although the research demonstrates that EEG alpha-biofeedback training per se is neither necessary for nor especially facilitative of the achievement of the alpha experience, the findings nevertheless add up to a very positive conclusion concerning self-regulation: The research demonstrates that we have greater abilities of self-control of experiential state than we have hitherto been willing to grant ourselves. As I concluded in an earlier article (Plotkin, 1976a),

> we should not be unduly disappointed that there is no direct association between enhanced alpha and the alpha experience. The chain of research on alpha feedback, from Kamiya's . . . first paper to the present, has been valuable in showing us that, although we once thought that a box of amplifiers and filters had made it possible to induce a de-

sirable state of consciousness more rapidly and effectively than ever before, in fact we were really always doing it "on our own." We simply discovered once again that often people only need a certain degree of faith in their natural powers and abilities, along with an appropriate setting and simple instructions, in order to accomplish what they feel is normally beyond their potential. . . . The power to enter altered states of consciousness is a *natural* ability that we all can potentially tap; learning how to do this without *external* devices such as electronics and drugs will serve to expand our behavior potential in the widest range of circumstances. (p. 97, italics in original)

Thus, it appears that Maslow (1969), for example, was somewhat mistaken when he concluded from the early alpha studies that "it is already possible to teach people how to feel happy and serene" (p. 728). It would now seem more appropriate to say—and, incidentally, this is more in keeping with humanistic themes—that we have discovered that it has always been possible for people to *allow themselves* to feel happy and serene.

A second related implication concerns the similarity between the alpha-feedback phenomenon and the hypnotic situation: Both are ways in which latent behavior potential can be evoked, and in a similar manner. One procedure for inducing the hypnotic state (Plotkin & Schwartz, Note 5) centers around the hypnotist's carefully timed redescriptions of behavior: The hypnotic subject's behavior is redescribed in such a way that the subject comes to see his or her own behavior as occurring "automatically" or under the "control" of the hypnotist. For example, the subject's arm may be seen as rising autonomously when the subject is, of course, the one who is actually raising it. Similarly, the biofeedback researcher gives an (unintentionally) inaccurate description of the alpha-feedback situation with the result that the trainee believes that the promised experiential state is a consequence of biofeedback-augmented alpha enhancement rather than the trainee's direct achievement (the latter being, in fact, the case). With this redescription of the trainees' behavior, the researcher has managed to circumvent the typical trainees' self-doubts about their abilities to put themselves in this state, which leaves them in a position in which they can simply go ahead and do just that (as long as they do not see it that way at the time). In essence, the trainees are supplied with a special description of the behavior whereby they self-induce a change in consciousness so that they do not fully recognize their behavior for what it is. Inasmuch as they think that their potential effectiveness is limited to controlling the feedback tone, they do not recognize the situation as one in which there is, in fact, a question about their ability to self-induce the alpha experience; or if they do recognize the situation, all the evidence is stacked against their self-doubts, since they are, after all, "objectively" succeeding at the task! In sum, although biofeedback researchers have not been fully aware of it, alpha-feedback training has been a situation in which biofeedback has been used as an element in a somewhat sophisticated social-influence process that can lead to the evocation of latent powers of self-control.

These conclusions should not be taken as a disparagement of biofeedback training; biofeedback does represent a valuable advance in our capacity to introduce ourselves to new realms of physiological and psychological self-control. Rather, this research suggests at least two cautions or reminders for biofeedback users and researchers. First, we must distinguish between the intrinsic and the instrumental uses of biofeedback training. Biofeedback is used for intrinsic purposes when the physiology that is being controlled, is being controlled for its own sake. For example, the use of biofeedback training for the reduction of high blood pressure or for muscular reeducation is an instance of intrinsic use: If the hypertensive can use the biofeedback monitor to learn to lower his blood pressure, or if the cerebral-palsy victim can employ the information supplied by an EMG monitor to learn to coordinate his movements once again, then there is no question that the physiological control itself is valuable. On the other hand, when biofeedback is used instrumentally, we cannot be as sure that the physiological control *per se* will be at all useful. Biofeedback is used instrumentally when the control of some aspect of our physiology is attempted not because this control is intrinsically valuable, but because it appears to

lead to—or to accompany—some other desirable state of affairs. For example, the use of EMG biofeedback training for anxiety reduction, or the use of EEG biofeedback training for "mind control," for altered-state induction, for pain control, or as psychotherapy for neurotics or alcoholics is an instrumental use. When biofeedback is used instrumentally there is a gap, usually a categorical gap, between the physiological process that is being self-regulated and the desired behavioral or psychological outcome. In such cases we must be most careful before concluding that the control of the physiology in question has any special relevance to the desired or attained goal.

The second reminder balances out the first: Biofeedback training is not merely a form of manipulation of human physiology; it is a complex social–behavioral interaction in which not merely physiology but attitudes, expectations, motivations, attention, experience, alertness, and understandings are being directly and indirectly influenced independently of any contingencies between physiology and feedback. The present article illustrates how biofeedback training can be more fully understood if it is viewed as a social-therapeutic activity with important physiological aspects, as opposed to being thought of as a strictly physiological training procedure with incidental (and perhaps annoying) social attributes. Especially when employed instrumentally, the general biofeedback framework is not merely a novel application of operant conditioning methodology but a potentially powerful context for the mobilization and activation of our latent self-control and self-healing capacities (Plotkin, Note 3). It is in this latter role that biofeedback training may find its most fruitful applications as a therapeutic tool.

Reference Notes

1. Plotkin, W. B. *Biofeedback-associated disinhibition of eyes-closed EEG alpha strength: Spontaneous or learned?* Manuscript submitted for publication, 1979.
2. Glaros, A. G. *Subjective reports in alpha feedback training*. Paper presented at the meeting of the Biofeedback Society of America, Orlando, Florida, March 1977.
3. Plotkin, W. B. *The placebo effect, self-healing, and biofeedback: The role of faith in therapy*. Manuscript submitted for publication, 1979.
4. Plotkin, W. B. *Individual differences in susceptibility to placebo-induced "alpha experiences."* Manuscript in preparation, 1979.
5. Plotkin, W. B., & Schwartz, W. R. *A conceptualization of hypnosis: Exploring the place of anomaly and appraisal in behavior and experience.* Manuscript submitted for publication, 1979.

References

Adrian, E. D., & Mathews, B. A. C. The Berger rhythm: Potential changes from the occipital lobes in man. *Brain*, 1934, *57*, 355–385.

Anand, B. K., Chhina, G. S., & Singh, B. Some aspects of electroencephalographic studies in yogis. *EEG and Clinical Neurophysiology*, 1961, *13*, 452–456.

Beatty, J. Similar effects of feedback signals and instructional information on EEG activity. *Physiology and Behavior*, 1972, *9*, 151–154.

Berger, H. Über das Elektrenkephalogramm des Menschen. *Journal für Psychologie und Neurologie*, 1930, *40*, 160–179.

Brown, B. Recognition of aspects of consciousness through association with EEG alpha activity represented by a light signal. *Psychophysiology*, 1970, *6*, 442–452.

Chatrian, G. E., Magnus, M. D., Petersen, C., & Lazarte, J. S. The blocking of rolandic wicket rhythm and some central changes related to movement. *EEG and Clinical Neurophysiology*, 1959, *11*, 497–510.

Davison, G. C., & Valins, S. Maintenance of self-attributed and drug-attributed behavior change. *Journal of Personality and Social Psychology*, 1969, *11*, 25–33.

DeGood, D. E., Elkin, B., Lessin, S., & Valle, R. S. Expectancy influence on self-reported experience during alpha feedback training: Subject and situational factors. *Biofeedback and Self-Regulation*, 1977, *2*, 183–194.

Deikman, A. J. Experimental meditation. *Journal of Nervous and Mental Disease*, 1963, *236*, 329–343.

Erickson, M. H., Rossi, E. L., & Rossi, S. *Hypnotic realities: The induction of clinical hypnosis and forms of indirect suggestion*. New York: Irvington, 1976.

Galin, D., & Ornstein, R. Lateral specialization of cognitive mode: An EEG study. *Psychophysiology*, 1972, *9*, 412–418.

Grossberg, J. M. Brain wave feedback experiments and the concept of mental mechanisms. *Journal of Behavior Therapy and Experimental Psychiatry*, 1972, *3*, 245–251.

Hardt, J. V., & Kamiya, J. Some comments on Plotkin's self-regulation of EEG alpha. *Journal of Experimental Psychology: General*, 1976, *105*, 100–108. (a)

Hardt, J. V., & Kamiya, J. Conflicting results in EEG alpha feedback studies: Why amplitude integration should replace percent time. *Biofeedback and Self-Regulation*, 1976, *1*, 63–75. (b)

Hart, J. Autocontrol of EEG alpha. *Psychophysiology*, 1968, *4*, 506. (Abstract)

Hilgard, E. R. Altered states of awareness. *Journal of Nervous and Mental Disease*, 1969, *149*, 68–79.

Hunt, H. T., & Chefurka, C. M. A test of the psychedelic model of altered states of consciousness: The role of introspective sensitization in eliciting unusual subjective reports. *Archives of General Psychiatry*, 1976, *33*, 867–876.

James, W. *The principles of psychology*. New York: Dover, 1950.

Jasper, H. H., & Penfield, W. Electrocorticograms in man: Effects of voluntary movement on the electrical activity of the precentral gyrus. *Archiv für Psychiatrie und Nervenkrankheiten*, 1949, *183*, 163–174.

Johnson, L. C. Learned control of brain wave activity. In J. Beatty & H. Legewie (Eds.), *Biofeedback and behavior*. New York: Plenum Press, 1977.

Kamiya, J. Conscious control of brain waves. *Psychology Today*, November 1968, *1*, 56–60.

Kamiya, J. Operant control of the EEG alpha rhythm and some of its reported effects on consciousness. In C. T. Tart (Ed.), *Altered states of consciousness*. New York: Wiley, 1969.

Kasamatsu, A., & Hirai, T. An electroencephalographic study of the Zen meditation (Zazen). In C. T. Tart (Ed.), *Altered states of consciousness*. New York: Wiley, 1969.

Klass, D. W., & Bickford, R. G. Observations on the rolandic arceau rhythm. *EEG and Clinical Neurophysiology*, 1957, *9*, 570.

Kreitman, N., & Shaw, J. C. Experimental enhancement of alpha activity. *EEG and Clinical Neurophysiology*, 1965, *18*, 147–155.

Lawrence, J. *Alpha brain waves*. New York: Avon, 1972.

Lindsley, D. B. Attention, consciousness, sleep, and wakefulness. In J. Field (Ed.), *Handbook of physiology: Section 1, Neurophysiology* (Vol. 1). Washington, D.C.: American Physiological Society, 1960.

Lynch, J. J., & Paskewitz, D. A. On the mechanisms of the feedback control of human brain wave activity. *Journal of Nervous and Mental Disease*, 1971, *153*, 205–217.

Lynch, J. J., Paskewitz, D. A., & Orne, M. T. Some factors in the feedback control of human alpha rhythm. *Psychosomatic Medicine*, 1974, *36*, 399–410.

Marshall, M. S., & Bentler, P. M. The effects of deep physical relaxation and low-frequency-alpha brainwaves on alpha subjective reports. *Psychophysiology*, 1976, *13*, 505–516.

Maslow, A. H. Toward a humanistic biology. *American Psychologist*, 1969, *24*, 724–735.

Morgan, A. H., MacDonald, H., & Hilgard, E. R. EEG alpha: Lateral asymmetry related to task and hypnotizability. *Psychophysiology*, 1974, *11*, 275–282.

Mulholland, T. B. Feedback electroencephalography. *Activas Nervosa Superior*, 1968, *10*, 410–438.

Mulholland, T. B. Occipital alpha revisited. *Psychological Bulletin*, 1972, *3*, 176–182.

Mulholland, T. B. Objective EEG methods for studying covert shifts of visual attention. In F. J. McGuigan & R. A. Schoonover (Eds.), *The psychophysiology of thinking: Studies of covert processes*. New York: Academic Press, 1973.

Mulholland, T. B., & Peper, E. Occipital alpha and accommodative vergence, pursuit tracking, and fast eye movements. *Psychophysiology*, 1971, *8*, 556–575.

Naranjo, C., & Ornstein, R. *On the psychology of meditation*. New York: Viking, 1971.

Nideffer, R. M. Alpha and the development of human potential. In D. Shapiro et al. (Eds.), *Biofeedback and self-control, 1972*. Chicago: Aldine-Atherton, 1973.

Nowlis, D. P., & Kamiya, J. The control of electroencephalographic alpha rhythms through auditory feedback and the associated mental activity. *Psychophysiology*, 1970, *6*, 476–484.

Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 1962, *17*, 776–783.

Orne, M. T., & Paskewitz, D. A. Aversive situational effects on alpha feedback training. *Science*, 1974, *186*, 458–460.

Ossorio, P. G. Never smile at a crocodile. *Journal for the Theory of Social Behavior*, 1973, *3*, 121–140.

Ossorio, P. G. *What actually happens: The representation of real world phenomena*. Columbia: University of South Carolina Press, 1978.

Paskewitz, D. A. EEG alpha activity and its relationship to altered states of consciousness. *Annals of the New York Academy of Sciences*, 1977, *296*, 154–161.

Paskewitz, D. A., Lynch, J. J., Orne, M. T., & Costello, J. The feedback control of alpha activity: Conditioning or disinhibition? *Psychophysiology*, 1970, *6*, 637–638. (Abstract)

Peek, C. J. A critical look at the theory of placebo. *Biofeedback and Self-Regulation*, 1977, *2*, 327–335.

Peper, E. Reduction of efferent motor commands during alpha feedback as a facilitator of EEG alpha and a precondition for changes in consciousness. *Kybernetic*, 1971, *9*, 226–231.

Peper, E. Localized EEG alpha feedback training: A possible technique for mapping subjective, conscious, and behavioral experiences. *Kybernetic*, 1972, *11*, 166–169.

Plotkin, W. B. On the self-regulation of the occipital alpha rhythm: Control strategies, states of consciousness, and the role of physiological feedback. *Journal of Experimental Psychology: General*, 1976, *105*, 66–69. (a)

Plotkin, W. B. Appraising the ephemeral "alpha phenomenon": A reply to Hardt and Kamiya. *Journal of Experimental Psychology: General*, 1976, *105*, 109–121. (b)

Plotkin, W. B. On the social psychology of experiential states associated with EEG alpha biofeedback training. In J. Beatty & H. Legewie (Eds.), *Biofeedback and behavior*. New York: Plenum Press, 1977.

Plotkin, W. B. Long-term eyes-closed alpha-enhancement training: Effects on alpha amplitudes and on experiential state. *Psychophysiology*, 1978, *15*, 40–52.

Plotkin, W. B. The role of attributions of responsibility in the facilitation of unusual experiential states during EEG alpha training: An analysis of the biofeedback-placebo effect. *Journal of Abnormal Psychology*, in press.

Plotkin, W. B., & Cohen, R. Occipital alpha and the attributes of the "alpha experience." *Psychophysiology*, 1976, *13*, 16–21.

Plotkin, W. B., Mazer, C., & Loewy, D. Alpha enhancement and the likelihood of an alpha experience. *Psychophysiology*, 1976, *13*, 466–471.

Regestein, Q. R., Pegram, V., Cook, B., & Bradley, D. Alpha rhythm percentage maintained during 4- and 12-hour feedback periods. In N. E. Miller et al. (Eds.), *Biofeedback and self-control, 1973*. Chicago: Aldine-Atherton, 1974.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, *80*(1, Whole No. 609).

Sacks, B., Fenwick, P. B. C., Marks, I., Fenton, G. W., & Hebden, A. An investigation of the phenomenon of autocontrol of the alpha rhythm and possible associated feeling states using visual feedback. *EEG and Clinical Neurophysiology*, 1972, *32*, 461.

Sarbin, T. R. Contextualism: A world view for modern psychology. In A. Landfield (Ed.), *Nebraska Symposium on Motivation* (Vol. 24). Lincoln: University of Nebraska Press, 1977.

Schwartz, G. E., Davidson, R. J., & Pugash, E. Voluntary control of patterns of EEG parietal asymmetry: Cognitive concomitants. *Psychophysiology*, 1976, *13*, 498–504.

Shapiro, D. Biofeedback and the regulation of complex psychological processes. In J. Beatty & H. Legewie (Eds.), *Biofeedback and behavior*. New York: Plenum Press, 1977.

Shor, R. E. Naturally occurring "hypnotic-like" experiences in the normal college population. *International Journal of Clinical and Experimental Hypnosis*, 1960, *8*, 151–163.

Spearman, C. *The nature of intelligence and the principles of cognition*. London: Macmillan, 1923.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Manual for the State–Trait Anxiety Inventory (Self-Evaluation Questionnaire)*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.

Stoyva, J., & Kamiya, J. Electrophysiological studies of dreaming as the prototype of a new strategy in the study of consciousness. *Psychological Review*, 1968, *75*, 192–205.

Strayer, F., Scott, W. B., & Bakan, P. A re-examination of alpha feedback training: Operant conditioning or perceptual differentiation? *Canadian Journal of Psychology*, 1973, *27*, 247–253.

Stroebel, C. F., & Glueck, B. C. Biofeedback treatment in medicine and psychiatry: An ultimate placebo? In L. Birk (Ed.), *Biofeedback: Behavioral Medicine*. New York: Grune & Stratton, 1973.

Titchener, E. B. Description vs. statement of meaning. *American Journal of Psychology*, 1912, *23*, 165–182.

Travis, T. A., Kondo, C. Y., & Knott, J. R. Subjective aspects of alpha enhancement. *British Journal of Psychiatry*, 1975, *127*, 122–126.

Valins, S., & Nisbett, R. E. Attribution processes in the development and treatment of emotional disorder. In E. E. Jones et al. (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, N.J.: General Learning Press, 1972.

Valle, R. S., & Levine, J. M. Expectation effects in alpha wave control. *Psychophysiology*, 1975, *12*, 306–309.

Wallace, R. K. Physiological effects of transcendental meditation. *Science*, 1970, *167*, 1751–1754.

Walsh, D. H. Interactive effects of alpha feedback and instructional set on subjective state. *Psychophysiology*, 1974, *11*, 428–435.

Wertheim, A. H. Oculomotor control and occipital alpha activity: A review and a hypothesis. *Acta Psychologica*, 1974, *38*, 235–256.

Zubeck, J. P. (Ed.). *Sensory deprivation: Fifteen years of research*. New York: Appleton-Century-Crofts, 1969.

Received May 30, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1149–1164

Statistical Adjustments and Uncontrolled Studies

Herbert I. Weisberg
The Huron Institute, Cambridge, Massachusetts

Many evaluations of social interventions are based on uncontrolled assignments of individuals to treatment groups. Statistical adjustments are often used to compensate for naturally occurring differences between groups. There is much confusion and controversy about the adequacy of these statistical methods. A variety of interrelated problems have been identified, including measurement error, unequal growth rates across groups, and regression artifacts. In this article it is shown that these problems can all be subsumed under a general conceptual framework, as particular examples of model misspecification. This perspective is helpful in revealing clearly the nature of the problems posed by lack of experimental control. The important case of linear adjustment (analysis of covariance) is given special attention. An expression is derived for the proportion of bias remaining after adjustment, in terms of easily interpretable parameters. Implications of these results for research and evaluation design are considered.

To evaluate the effectiveness of a social intervention, the performance of a group receiving the "treatment" must be compared with a standard representing the expected performance in the absence of intervention. The fundamental problem in research design is to find a valid standard of comparison. Randomization is generally accepted as the ideal approach. That is, we use a random mechanism to assign individuals to either a treatment group or an untreated control group. Random selection virtually guarantees (at least for large samples) that the control group's performance will correspond to that of the treatment group without the intervention. So a straightforward comparison of mean outcomes for the two groups will provide an unbiased estimate of the treatment's effect.

Often, however, it is impossible to exercise experimental control. Complex social forces unknown to the investigator determine which individuals wind up in each of the groups. With such uncontrolled selection designs, the straightforward difference of group means may be a biased estimate of the effect. In these situations a variety of statistical methods have been proposed to compensate for this bias and thus provide an unbiased estimate. The analysis of covariance (ANCOVA) is perhaps most widely used for this purpose.

Recently there has been a great deal of concern about the adequacy of ANCOVA and other statistical adjustment procedures. Several investigators have shown that under models representing uncontrolled selection, the ANCOVA may either overadjust or underadjust (Bryk & Weisberg, 1977; Cain, 1975; Cochran & Rubin, 1973; Cronbach, Rogosa, Floden, & Price, Note 1). The estimates generated may in some instances be seriously misleading. It is even possible for the remaining bias after adjustment to be larger in absolute value than the initial bias without any adjustment.

Confusion over the adequacy of statistical adjustments is part of a larger debate about the usefulness of designs based on uncontrolled

---

This work was supported by Grant NIE–G–76–0090 from the National Institute of Education, U.S. Department of Health, Education and Welfare. However, points of view or opinions stated do not represent official NIE position or policy.

The author gratefully acknowledges the contribution of Anthony S. Bryk to the development of the ideas expressed in this article and his many helpful comments on earlier drafts. Sincere thanks also to Walt Haney, David Rogosa, and the referees for their valuable suggestions, many of which have been incorporated in the final version.

Requests for reprints should be sent to Herbert I. Weisberg, The Huron Institute, 123 Mount Auburn Street, Cambridge, Massachusetts 02138.

studies. Some analysts (see Campbell & Boruch, 1975; Gilbert, Light, & Mosteller, 1975; Riecken & Boruch, 1974) argue forcefully that randomization is essential and often feasible. Yet the vast majority of social research is based on uncontrolled selection, because a wide variety of practical, ethical, and political problems prevent the implementation of rigorous randomized experiments (e.g., see Cohen, 1975; Suchman, 1967; Weiss, 1972). Thus it is crucially important to understand exactly what problems are associated with uncontrolled studies.

The present article proposes a conceptual framework that may help to clarify these issues. I shall begin with a discussion of statistical adjustment in general, and then consider in detail the important special case of linear models. The perspective is similar to that of several other investigators (Barnow & Cain, 1977; Cain, 1975; Cochran & Rubin, 1973; Cronbach et al., Note 1; Goldberger, Note 2), and some of the specific results can be found in their work. One of the main objectives of this review is to draw together insights scattered throughout the extensive literature on the adequacy of adjustments.

Statistical Adjustments for Confounding Variables

To avoid confusion, we will not consider the effects of finite sample sizes, or the accompanying problems of estimating parameters on the basis of data. Such issues compound the problems addressed here but do not affect the basic argument. In effect, we will be considering the case of large samples, so that the precision of estimated parameters is very high.

Let us consider the study sample as randomly drawn from some population. Individuals in the sample are assigned to a treatment or control group on the basis of a mechanism that may or may not be known. Following Rubin (1974), we can think of two potential outcomes corresponding to each individual: the outcome received under the treatment and that received without any intervention. We can define Y_i: observed outcome under treatment actually received, W_i: outcome that would have been observed under treatment, and Z_i: outcome that would have been observed under control conditions.

These definitions may at first seem confusing. To understand them clearly, it may be helpful for the reader to imagine a two-stage process consisting of group selection and treatment administration. In the first stage, individuals are assigned to the treatment and control groups as they would be at the beginning of a study. At the second stage, however, there are three alternative possibilities.

First, the study may be carried out as planned, with subjects who are assigned to the treatment group receiving the treatment and control subjects experiencing no intervention. In this case, the observed outcome corresponds to the variable Y defined above. The second possibility is to suspend the study and simply give the treatment to all subjects. The outcome in this case would be W. Finally, we can suspend the study but give the treatment to no one. Then the outcome would be Z.

Ideally, the effect of the treatment would be assessed by comparing W_i and Z_i for each individual i. But of course, in general, the second and third options described above are strictly hypothetical, and we will have information only on Y. The key question in uncontrolled studies is whether we can obtain a useful and valid estimate of effect when only Y is available. Under what circumstances will an actual study allow us to make inferences about what would have occurred under different scenarios?

A general answer to this question would be extremely complex. Note that for each individual the treatment effect is given by

$$\alpha_i = W_i - Z_i. \tag{1}$$

In general, α_i may depend on individual characteristics and even on the selection mechanism. For example, suppose we happen to assign to the treatment those who can benefit from it the most. Then the average effect for those in the treatment group will be relatively large. But the results will generalize only to similarly selected groups. When the α_is are related to individual characteristics in this way, caution is required in the interpretation of average effects.

Although this issue of interactive effects is an important and confusing one, I wish in this article to highlight the problems pertaining specifically to statistical adjustments. Let us

therefore restrict consideration to the situation in which the treatment effect is constant across individuals. That is,

$$W_i = Z_i + \alpha, \tag{2}$$

so that we can write

$$Y_i = Z_i + Q\alpha, \tag{3}$$

where $Q_i = 1$ if subject i is in the treatment group and $Q_i = 0$ if subject i is in the control group. Note that if Q is determined by a random mechanism, then each individual has the same probability of being assigned to the treatment group. Let P be the total proportion assigned to the treatment. Then

$$P(Q = 1) = P$$

and

$$P(Q = 0) = 1 - P. \tag{4}$$

Under randomization, the expected value of Z for those in each group should be identical. That is, if we define

$$\mu_{Z1} = E(Z|Q = 1)$$

and

$$\mu_{Z0} = E(Z|Q = 0), \tag{5}$$

we would have

$$\mu_{Z1} = \mu_{Z0}. \tag{6}$$

If, however, the assignment mechanism is nonrandom, it is possible that $\mu_{Z1} \neq \mu_{Z0}$. Even with no intervention, there might be a difference in mean outcomes between the groups. In this case, if we simply compare the mean observed outcomes for the groups, we will obtain on the average (and approximately for large sample sizes)

$$\mu_{Y1} - \mu_{Y0} = \alpha + \mu_{Z1} - \mu_{Z0}. \tag{7}$$

That is, the estimated effect will be inflated by $\mu_{Z1} - \mu_{Z0}$. While under randomization $\mu_{Z1} - \mu_{Z0}$ will be 0, under nonrandom assignment it may be either positive or negative. This term represents the *selection bias* in estimating α. Equations 5 and 7 show that the bias is determined by the relationship between Z and Q in the study sample. In fact, it will be convenient for us to reexpress this relationship, using the well-known formula for the point biserial correlation (e.g., see McNemar, 1969, p. 218) as:

$$\mu_{Z1} - \mu_{Z0} = \frac{\sigma_Z}{\sqrt{P(1-P)}}\rho_{ZQ}, \tag{8}$$

where σ_Z is the overall standard deviation of Z and ρ_{ZQ} is the correlation between Z and Q.

The dilemma posed by uncontrolled studies is that this correlation cannot be estimated empirically. Without the intervention, the distribution of Z could be obtained, but there would be no information on Q. After the study, Q can be observed, but Z is no longer observable. The observable outcome is then the modified variable Y that includes a component attributable to the treatment. From Equation 3 it is clear that the relationship between Y and Q is not the same as that between Z and Q. Thus the crucial relationship between Z and Q cannot be observed.

In many situations, however, it is possible to identify other variables that are related to the assignment variable Q, and which are thought to "explain" the relationship between Z and Q. That is, differences between μ_{Z1} and μ_{Z0} are caused (at least in part) by differences between groups in the distribution of these "confounding factors." If such factors can be identified and measured, it may be possible to compensate for their effects in estimating α. Statistical adjustments are based on this idea.

Intuitively, a confounding factor can be defined as a variable that has a different distribution in each of the treatment groups and that is causally related to the outcome variable. This definition is ambiguous, however, because the concept of causality is difficult to operationalize. Whatever we mean by causal influence, it seems likely that in general the value of $\mu_{Z1} - \mu_{Z0}$ will be determined by a complex combination of variables. The actual selection process, moreover, may be described in more than one way. We may have many equally plausible "explanations" for group differences. So it may be unrealistic to expect an unambiguous criterion defining a confounding factor.

On the other hand, it is not unreasonable to ask whether in a particular study, using statistical adjustment and a given set of adjustment variables (covariates), an unbiased estimate of α can be expected. Further, we can try to specify the general conditions under which adjustment will be useful. What properties must X have in order to eliminate bias completely, or at least reduce it to an acceptable level? Can we assess the amount of bias

remaining after adjustment? These questions will be addressed in the remainder of this article.

Complete Elimination of Bias

Let $E(Z|X)$ be the conditional expectation, or mean, of Z given X. It is the average value of Z in the study population for individuals having a given value of X. Thus it is in general a mathematical function of X.

Similarly, let $E(Z|X, 1)$ be the conditional expectation of Z given X for those individuals assigned to the treatment group. Let $E(Z|X, 0)$ be the conditional expectation for control group subjects. In general, these expectations will not be equal for a given value of X. Suppose, however, that

$$E(Z|X, 1) = E(Z|X, 0) \qquad (9)$$

or equivalently that

$$E(Z|X, Q) = E(Z|X). \qquad (10)$$

This means that even though the unconditional expectations for the two groups differ, the conditional means are equal. Conditioning on X eliminates the selection bias. We will say in this case that X is a *complete* confounding factor.

To see more clearly the importance of this condition, suppose we "match" two individuals on the basis of a variable X for which Equation 10 holds. Let Z_1 be the outcome for the treatment group subject and Z_0 be the outcome for the control group subject. Then

$$E(Z_1|X, 1) - E(Z_0|X, 0) = 0, \qquad (11)$$

so that

$$E(Y_1|X, 1) - E(Y_0|X, 0) = \alpha. \qquad (12)$$

By matching individuals on the basis of a complete confounding variable, we can obtain an unbiased estimate of α. Such matching procedures are commonly employed, particularly in medical research. There are of course many practical problems to deal with, even though the method is theoretically correct. Rubin (1973a, 1973b) provides an excellent discussion of these problems. A less technical but more comprehensive review is provided in Anderson et al. (in press).

Note that there may be many different complete confounding factors satisfying Equation 10. If they vary in terms of the strength of their relationship with Z, some factors will result in more precise estimates of α than others will. In terms of bias reduction, however, they are all equivalent. Note also that X may be a vector consisting of several variables. Equation 10 is still the defining property.

Now suppose that a "covariate" X (either univariate or multivariate) is used for adjustment, and Equation 10 does not hold. Then there may be differences between groups that are unrelated to X. So after adjustment for X, there may remain an apparent treatment effect really attributable to preexisting group differences.

The controversy over statistical adjustments mentioned above centers on the adequacy of the Xs used in practical applications. If adjustment proceeds as if Equation 10 holds, and a more general model of the form

$$E(Z|X, Q) = f(X, Q) \qquad (13)$$

actually holds, how well do these procedures perform?

Stated in this way, the adequacy of statistical adjustment may be seen as a problem of model specification. Essentially, we are estimating a parameter α under a particular restricted form of the relationship between Z, X, and Q. We wish to know how accurate our estimate will be if a more general model actually holds.

Note that from this perspective, the important issues concern the relationships among real, *potentially* observable variables. Although it is true that Z and Q cannot be observed together in the same study, each could be observed if we were willing to forego the other. The adequacy of adjustment depends on how the covariate X relates to these two variables.

It is perhaps worth contrasting this perspective with that adopted in the widely circulated paper by Cronbach and his Stanford colleagues (Cronbach et al., Note 1). These authors define a true model in terms of two ideal variables. The performance of adjustments is then assessed in terms of the relationship between the covariates actually used and these unknown, ideal factors. However, the ambiguity inherent in defining and interpreting these constructs and the complexity of the

mathematical analysis have led to some confusion. By avoiding hypothetical variables, I hope to provide a clearer understanding of the problems encountered in practical situations.

Note that while Equation 10 specifies a theoretical condition for complete adjustment, this condition is difficult to verify in practice. Generally speaking, we can be sure that the condition is fulfilled only when the investigator has complete control over the assignment of subjects to treatment groups. With complete control, the experimenter has two main options: randomization and explicit control.

When individuals can be assigned randomly to the two groups, $\mu_{Z1} - \mu_{Z0} = 0$, and statistical adjustment is unnecessary. If adjustment is performed using any covariate, the estimate of effect remains unbiased and may have greater precision. Increase of precision in randomized experiments was in fact the original purpose of the ANCOVA (Fisher, 1932).

By explicit control, we mean that a covariate X serves as the sole basis for group assignment. That is, the probability of being in each group can depend on X but no other variable. In practice, this amounts to conditional randomization, since for a given value of X, individuals are assigned randomly with the value of P a function of X. Rubin (1977) has recently advocated designs based on explicit control for educational research.

In the extreme case when the probability of assignment to the treatment group is either 0 or 1 for any given X value, we have deterministic assignment conditional on X. The special case when all individuals below a certain cutoff score on X are assigned to one group and those above to the other group has been called the *regression discontinuity design* (Campbell, 1969; Campbell & Stanley, 1966). The problem with this design is that, although X is indeed a complete confounding variable, the distribution of X in the two groups does not overlap. Thus we cannot match individuals on the basis of X, and we must rely heavily on model assumptions in order to analyze such experiments.

As mentioned above, both randomization and explicit control require that the investigator be able to determine the assignment procedure. In fact, these two techniques constitute the backbone of controlled experimentation in the Fisherian tradition, and a huge literature has developed to elaborate on these two basic ideas (e.g., Cochran & Cox, 1957; Cox, 1958; Kempthorne, 1952; Winer, 1971). This tradition relies on experimental control as the prerequisite for causal inferences.

In our terms, then, experimental control is the only general method of insuring that the covariates employed in a given study constitute complete confounding factors. With uncontrolled studies, we will not know how much bias remains after adjustment. Moreover, because there is an unlimited class of possible models relating Z, X, and Q, we cannot hope to answer this question definitively. Restricting consideration to linear models, however, will provide some helpful insight.

Linear Models and the Analysis of Covariance

Let us assume that

$$E(Z|X, Q) = \mu + \beta X + \delta Q. \qquad (14)$$

In this model δ can be interpreted as the expected difference in Z for individuals in the two groups with identical X values. Now if X is a complete covariate, then Equation 14 reduces to

$$E(Z|X, Q) = \mu + \beta X, \qquad (15)$$

so that

$$\mu_{Z1} - \mu_{Z0} = \beta(\mu_{X1} - \mu_{X0}), \qquad (16)$$

and we could form the estimator

$$\hat{\alpha} = \mu_{Y1} - \mu_{Y0} - \beta(\mu_{X1} - \mu_{X0}). \qquad (17)$$

From Equations 7 and 16, we see that $\hat{\alpha}$ will be an unbiased estimator of α.

If we ignore finite-sample estimation problems, Equation 17 represents the ANCOVA estimate of α. Now suppose that an ANCOVA is employed when the true underlying model is given by Equation 14. Note first that δ represents the bias remaining after adjustment by ANCOVA with X as the covariate. We will be interested in the relationship between δ and the initial bias without adjustment $(\mu_{Z1} - \mu_{Z0})$. The ratio of these quantities may be termed the proportion of initial bias remaining, which we will denote π. We use the term proportion

in a general sense, since π need not necessarily lie between 0 and 1.

Note that from Equations 3 and 14 we have

$$E(Y|X, Q) = \mu + \beta X + (\delta + \alpha)Q. \quad (18)$$

So α and δ are totally confounded in terms of the linear model relating observable variables. Thus instead of estimating α as we could if X were complete, we can only estimate $\delta + \alpha$.

Let us denote the matrix of correlations among Z, X, and Q by

$$\begin{bmatrix} 1 & \rho_{ZX} & \rho_{ZQ} \\ \rho_{ZX} & 1 & \rho_{XQ} \\ \rho_{ZQ} & \rho_{XQ} & 1 \end{bmatrix}.$$

Consider the basic model specified by Equation 15. Using standard results we can write

$$\delta = \rho_{ZQ\cdot X}\frac{\sigma_{Z\cdot X}}{\sigma_{Q\cdot X}}, \quad (19)$$

where

$\rho_{ZQ\cdot X}$ = partial correlation of Z and Q given X,
$\sigma_{Z\cdot X}^2$ = conditional variance of Z given X,

and

$\sigma_{Q\cdot X}^2$ = conditional variance of Q given X.

Further we have

$$\sigma_{Z\cdot X}^2 = \sigma_Z^2(1 - \rho_{ZX}^2) \quad (20)$$

and

$$\sigma_{Q\cdot X}^2 = \sigma_Q^2(1 - \rho_{XQ}^2) = P(1 - P)(1 - \rho_{XQ}^2). \quad (21)$$

From Equations 19, 20, and 21,

$$\delta = \rho_{ZQ\cdot X}\frac{\sigma_Z}{\sqrt{P(1 - P)}}\frac{\sqrt{1 - \rho_{ZX}^2}}{\sqrt{1 - \rho_{XQ}^2}}. \quad (22)$$

Now, using Equation 8, we can express the proportion of bias remaining after adjustment as

$$\pi = \frac{\rho_{ZQ\cdot X}}{\rho_{ZQ}}\frac{\sqrt{1 - \rho_{ZX}^2}}{\sqrt{1 - \rho_{XQ}^2}}. \quad (23)$$

Note that π may be either positive (underadjustment) or negative (overadjustment), depending on the signs of $\rho_{ZQ\cdot X}$ and ρ_{ZQ}. Moreover, $\rho_{ZQ\cdot X} = 0$ is a necessary and sufficient condition for ANCOVA to be unbiased under the model we are considering. That is, the correlation between Z and Q must be reduced to zero by conditioning on X.

Table 1
Range of π for Different Combinations of ρ_{ZQ}, ρ_{XZ}, ρ_{ZX}

Basic situation	Case	Sign (ρ_{ZQ})	Sign (ρ_{XQ})	Sign (ρ_{ZX})	π
1	1	+	+	+	$-\infty$ to +1
	2	+	−	−	$-\infty$ to +1
	3	−	+	−	$-\infty$ to +1
	4	−	−	+	$-\infty$ to +1
2	5	−	−	−	1 to $+\infty$
	6	−	+	+	1 to $+\infty$
	7	+	−	+	1 to $+\infty$
	8	+	+	−	1 to $+\infty$

Although this characterization of the condition for unbiased estimation is intuitively appealing, it is not very helpful for specifying when various π values are likely to occur in practice. However, using the definition of partial correlation, Equation 23 can be rewritten as

$$\pi = \frac{\rho_{ZQ} - \rho_{ZX}\rho_{XQ}}{\rho_{ZQ}(1 - \rho_{XQ}^2)}. \quad (24)$$

That is, we can express the proportion of bias remaining after adjustment as a function of the simple correlations among Z, X, and Q.

Suppose first that all three of these correlations are positive. Then it can be shown that $\pi < 1$, which means the initial bias will be reduced. Moreover, if ρ_{XQ} is large relative to ρ_{ZQ}, π could be very large in the negative direction. This means the adjustment can overcompensate for the initial bias. The range of possible values is $(-\infty, +1)$. A similar analysis can be undertaken for each combination of signs of the correlations. The results are presented in Table 1.

The eight cases can be classified into two basic situations. In Basic Situation 1 (Cases 1–4), the relationships are such that adjustment is in the right direction, reducing the initial bias. In Basic Situation 2 (Cases 5–8), adjustment will be in the wrong direction, further inflating the initial bias. The reason why each of the four cases in each basic situation has the same range of possible π values is that the four cases are really equivalent. Consider Case 1. Suppose we replace X by $-X$ in any Case 1 situation. We would not expect the amount of bias removed to depend

on whether X or $-X$ is used as a covariate, since both variables carry the same information. But if Z, X, and Q satisfy the conditions of Case 1, then Z, $-X$, and Q satisfy the conditions of Case 2. So each Case 2 situation may be viewed as a Case 1 situation with X replaced by $-X$. Similarly Case 3 can be generated by substituting $-Z$ for Z in Case 1 situations, and Case 4 by substituting $1 - Q$ for Q. Thus we need consider only the two basic situations.

Note that under Basic Situation 1, the initial bias increases, resulting in an estimate farther from the true value than the unadjusted mean difference. Although such bias inflation can in theory occur, it can probably be avoided in most practical situations. Often, enough is known about the general nature of selection to indicate at least the signs of the correlations among Z, X, and Q. For example, in the evaluation of compensatory education programs like Head Start (Campbell & Erlebacher, 1970), a very disadvantaged treatment group is compared with a somewhat less disadvantaged control group. For the outcomes and covariates commonly used, it can be expected that both ρ_{ZQ} and ρ_{XQ} will be negative and ρ_{ZX} positive. Under these conditions, we can be sure that Basic Situation 1 obtains, meaning that ANCOVA will underestimate the actual treatment effect.

For most practical applications of ANCOVA, Basic Situation 1 can be expected to hold. So it is of interest to consider in detail the expression for π in this case. In particular, is it possible in actual situations to say more about the range of possible π values? Because all four cases within Basic Situation 1 are equivalent in the sense described above, it suffices to consider only one case. We shall therefore assume that ρ_{ZQ}, ρ_{XQ}, and ρ_{ZX} are all positive.

Of the three correlations, only ρ_{XQ} can be estimated directly from the data. Note, however, that under our assumption of a constant treatment effect,

$$\rho_{YX\cdot Q} = \rho_{ZX\cdot Q}. \tag{25}$$

That is, $\rho_{ZX\cdot Q}$ can be estimated from the within-group relationship between Y and X, and can therefore be estimated from the data.

Using the standard formula for partial correlation

$$\rho_{ZX\cdot Q} = \frac{\rho_{ZX} - \rho_{ZQ}\rho_{XQ}}{\sqrt{(1-\rho_{ZQ}^2)(1-\rho_{XQ}^2)}}, \tag{26}$$

we obtain

$$\rho_{ZX} = \rho_{ZQ}\rho_{XQ} + \rho_{ZX\cdot Q}\sqrt{(1-\rho_{ZQ}^2)(1-\rho_{XQ}^2)}, \tag{27}$$

where the radical represents the *positive* square root. Substituting in Equation 24 yields

$$\pi = 1 - \rho_{ZX\cdot Q}\frac{\rho_{XQ}}{\rho_{ZQ}}\frac{\sqrt{1-\rho_{ZQ}^2}}{\sqrt{1-\rho_{XQ}^2}}. \tag{28}$$

So even if we can estimate ρ_{XQ} and $\rho_{ZX\cdot Q}$, we still require information on ρ_{ZQ} in order to assess π. Moreover, for fixed values of ρ_{XQ} and $\rho_{ZX\cdot Q}$, the value of π is quite sensitive to the value of ρ_{ZQ}. Since there is no constraint on ρ_{ZQ} for given values of ρ_{XQ} and $\rho_{ZX\cdot Q}$ (it can take any value between 0 and 1), we cannot place useful bounds on π in any obvious way. Only if some additional constraint on the correlations can be assumed will it be possible to restrict π to a subinterval of $(-\infty, 1)$.

Equation 28 provides another characterization of the condition for complete adjustment. All bias will be removed only if

$$\rho_{ZX\cdot Q} = \frac{\rho_{ZQ}\sqrt{1-\rho_{XQ}^2}}{\rho_{XQ}\sqrt{1-\rho_{ZQ}^2}}. \tag{29}$$

If $\rho_{ZX\cdot Q}$ is too small *or too large* there will be nonzero bias. The fact that the within-group correlation between Z and X could be too large under certain circumstances may at first appear counterintuitive. In particular, it seems plausible that ANCOVA would be unbiased when $\rho_{ZX\cdot Q} = 1$. But this will be true only if $\rho_{ZQ} = \rho_{XQ}$, or equivalently if $\rho_{ZX} = 1$. If the overall correlation between Z and X is less than 1, but the selection process results in a within-group correlation equal to 1, then adjustment will not be complete.

Equation 28 allows an intuitive understanding of the adjustment problem. If $\rho_{ZQ} > \rho_{XQ}$, the outcome (in the absence of intervention) is more strongly related to assignment than is the covariate. This means that the adjustment coefficient would have to be large in order to adjust fully, larger than β used in the analysis of covariance. If, on the other

hand, $\rho_{ZQ} < \rho_{XQ}$, then a modest adjustment coefficient is needed, and β may be either too small or too large.

In the methodological literature, a variety of potential problems in using statistical adjustment have been pointed out. I believe that all of these can be understood clearly in terms of the framework presented above. In the remainder of this article, I discuss some special issues that have received a great deal of attention. First we shall consider the problems raised by alternative models for individual growth over time on the outcome dimension. The important special case known as Lord's paradox (Lord, 1967) is analyzed in detail. Then we shall examine the situation in which covariates are measured with error, and finally we shall consider so-called regression effects.

Growth Models

Suppose the outcome of interest consists of the level of growth attained by an individual on some important dimension. Often the selection mechanism will result in a mean growth rate for the treatment group that is higher or lower than that of the controls, even in the absence of an intervention. Generally, in this situation, the covariate used is a pretest measured on the same dimension as the outcome score (posttest). In our notation, then, Y = observed posttest score, Z = posttest that would be observed without intervention, and X = pretest score. The use of statistical adjustments in this situation can be related to the voluminous literature on the measurement of change (see Cronbach & Furby, 1970), based on traditional psychometric assumptions. The problems in measuring change, from this perspective, depend primarily on reliability considerations. Much recent research on statistical adjustments when individuals are growing (Campbell & Boruch, 1975; Campbell & Erlebacher, 1970; Kenny, 1975) comes from this psychometric tradition. As a result, the problems caused by fallible measurement have been confounded with those related to growth per se. I shall discuss measurement error separately in a later section. My purpose here is to consider the problems caused by differential growth across treatment groups, even when pretests and posttests are perfectly reliable.

My analysis will follow closely that of Bryk and Weisberg (1977). They have considered various models for individual growth and for the selection of individuals into groups. With these models, they have identified situations in which various adjustments (including ANCOVA) will overadjust or underadjust.

I shall restrict consideration here to the special case when Equation 14 holds. Furthermore, it will be helpful to define standardized group differences on pretest and posttest by

$$D_Z = \frac{\mu_{Z1} - \mu_{Z0}}{\sigma_{Z\cdot Q}} \tag{30}$$

and

$$D_X = \frac{\mu_{X1} - \mu_{X0}}{\sigma_{X\cdot Q}}.$$

Then there are four basic ways that the relationship between the groups can change over time. First, the standardized distance between groups can remain the same over time. We refer to this situation as *standardized parallel growth*. Second, the standardized distance can increase. We call this situation *standardized divergence*. Third, the means may *cross over* between pretest and posttest, with the group that is higher on the pretest being lower on the posttest. Fourth, the standardized difference may decrease but not so much that crossover occurs. We then have *standardized convergence*. These four cases are illustrated in Figure 1 and summarized as follows:

I. Standardized parallel growth: $D_X = D_Z > 0.$

II. Standardized divergence: $0 < D_X < D_Z.$

III. Crossover: $D_X > 0;\quad D_Z < 0.$

IV. Standardized convergence: $D_X > D_Z > 0.$

Let us consider now what happens when the analysis of covariance is applied in each of the four cases. It will be convenient to rewrite Equation 28 in still another form:

$$\pi = 1 - \rho_{ZX\cdot Q}\frac{D_X}{D_Z}. \tag{31}$$

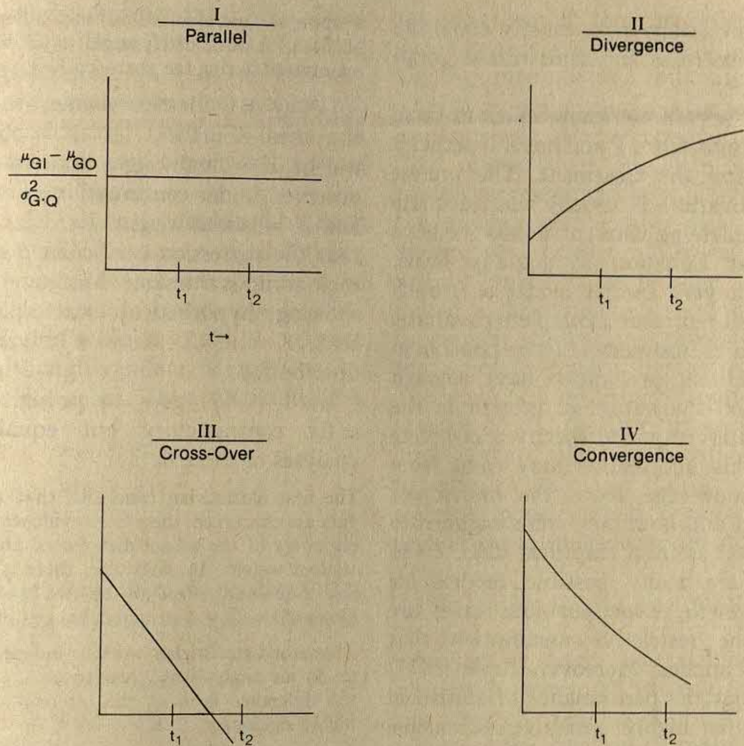

Figure 1. Types of standardized growth. [G(t) = growth at time t; t_1 = pretest time; t_2 = posttest time.]

Consider first the parallel case. From Equation 31 we obtain

$$\pi = 1 - \rho_{ZX \cdot Q}. \tag{32}$$

So in general, ANCOVA will underadjust under a parallel growth model. The special case when $\rho_{ZX \cdot Q} = 1$ and $D_X = D_Z$ is particularly interesting. It can be shown to be consistent with a "degenerate fan spread" model (Bryk & Weisberg, 1977). Under this model individual growth is represented as a straight line, and the slopes of these lines may vary across individuals, but nonnegligible growth for each begins at the same time, τ. This special situation is illustrated in Figure 2. As Bryk and Weisberg have shown, any reasonable technique will correctly adjust in this case. Because the pretest and the slope are perfectly correlated, the pretest contains full information on growth in the absence of intervention. So it constitutes a complete covariate in the sense defined above.

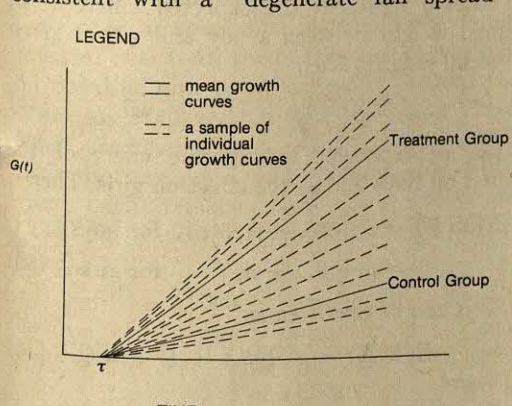

Figure 2. Degenerate fan spread. [G(t) = growth at time t.]

With a standardized divergence situation, π must be less than 1; so ANCOVA will underadjust. Under a crossover model, π will be greater than 1, and the initial bias becomes inflated. Only under the standardized convergence model does ANCOVA have the possibility of being correct. This will occur when

$$\rho_{ZX \cdot Q} = \frac{D_Z}{D_X}. \tag{33}$$

That is, the correlation between pretest and

posttest within groups must exactly equal the proportional decrease in standardized mean difference.

In general, we will not know which of these situations obtains, since Y and not Z is actually observable after the treatment. The pretest alone as a covariate is usually not adequate to allow complete adjustment unless we have evidence that Equation 33 actually holds. Only under a very special model of growth will the pretest represent a complete covariate.

If statistical adjustment is to be possible in growth situations, we must have enough information on the nature of growth in the absence of intervention to specify a complete covariate. This information may come from theoretical knowledge about the individuals being studied and how they were assigned to groups, as well as from empirical data. However, there are many possible models for individual growth. Even our four cases are all under the restrictive assumption that Equation 14 applies. Moreover, Bryk (1977) has shown that the performance of statistical adjustments is highly sensitive to model assumptions. Thus the use of statistical adjustments to control for differential growth appears quite problematic.

Lord's Paradox

Lord (1967) has presented a particularly interesting example of the problems in model specification when individuals are growing. He considered the following situation:

> A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight in the following June are recorded.
>
> At the end of the school year, the data are independently examined by two statisticians. Both statisticians divide the students according to sex. The first statistician examines the mean weight of the girls at the beginning of the year and at the end of the year and finds these to be identical. On further investigation, he finds that the frequency distribution of weight for the girls at the end of the year is actually the same as it was at the beginning.
>
> He finds the same to be true for the boys. Although the weight of individual boys and girls has usually changed during the course of the year, perhaps by a considerable amount, the group of girls considered as a whole has not changed in weight, nor has the group of boys. A sort of dynamic equilibrium has been maintained during the year. (p. 304)

Let $Q = 0$ for the women and $Q = 1$ for the men. Further, let Y = observed final weight, Z = final weight that would have been observed under continuation of previous diet, and X = initial weight. Lord further assumes that the regression coefficient β of Z on X in each group is the same. Moreover, his diagram showing hypothetical scatterplots suggests that Y and X have a bivariate normal distribution.

Lord (1967) goes to present two apparently contradictory but equally plausible analyses of these data:

> The first statistician concludes that as far as these data are concerned, there is no evidence of any interesting effect of the school diet (or of anything else) on student weight. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change.
>
> The second statistician, working independently, decides to do an analysis of covariance . . .. He finds that the difference between the intercepts is statistically highly significant.
>
> The second statistician concludes as is customary in such cases, that the boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes. (p. 305)

Lord concluded that there is no way to tell from the data available which results should be accepted. He infers that in general with uncontrolled studies, there is no way to make proper allowance for differences between treatment groups.

Let δ represent the expected difference in final weight between a boy and girl of given initial weight. Then

$$E(Z|X) = \mu + \beta X + \delta Q. \tag{34}$$

Let α_1 represent the specific effect of the diet on boys and α_2 the effect on girls. Then

$$\begin{aligned} E(Y|X) &= \mu + \beta X + \delta + \alpha_1 \quad \text{for boys} \\ &= \mu + \beta X + \alpha_2 \qquad\quad \text{for girls.} \end{aligned} \tag{35}$$

This can be rewritten as

$$E(Y|X) = \mu' + \beta X + (\delta + \alpha)Q, \tag{36}$$

where

$$\mu' = \mu + \alpha_2$$

and

$$\alpha = \alpha_1 - \alpha_2. \tag{37}$$

The general model representing Lord's situation has the form of Equation 18. So estimating the differential effect $\alpha_1 - \alpha_2$ is formally identical to the usual estimation problem we have been considering. Lord's paradox can be viewed in terms of model specification along the lines developed above.

As we shall see, the data described by Lord are consistent with a variety of underlying models. One possible model corresponds to the assumption that without the change in diet, the distribution of weights would be constant over time, although random fluctuations for individuals might occur. Then we would have

$$\mu_{Z0} = \mu_{X0},$$

$$\mu_{Z1} = \mu_{X1},$$

and

$$\sigma_{Z \cdot Q}^2 = \sigma_{X \cdot Q}^2. \tag{38}$$

Suppose further that there is no treatment effect for either sex ($\alpha_1 = \alpha_2 = 0$). Then we would observe

$$\mu_{Y1} = \mu_{X1},$$

$$\mu_{Y0} = \mu_{X0},$$

and

$$\sigma_{Y \cdot Q}^2 = \sigma_{X \cdot Q}^2, \tag{39}$$

which corresponds with Lord's description of data. This model represents the first statistician's construction of the situation.

Note that under this model, $D_Z = D_X$, and we would have standardized parallel growth. Moreover, since we know that the within-group correlation between Z and X is less than 1, it follows from Equation 32 that ANCOVA will underadjust.

Suppose, however, that instead of remaining constant over time, the distribution of weights for boys and girls would change in the natural course of events, even with the same diet. There are several forms such a change could take. For example, the variance for each group might remain constant, but the mean levels move closer together. Thus we might have

$$\mu_{Z1} < \mu_{X1},$$

$$\mu_{Z0} > \mu_{X0}, \tag{40}$$

and

$$\sigma_{Z \cdot Q}^2 = \sigma_{X \cdot Q}^2.$$

This implies that $D_Z < D_X$, and we have a standardized fan close situation. From Equation 33 then, if the correlation between Z and X within groups happens to be equal to D_Z/D_X, then ANCOVA will adjust perfectly. This situation must hold if the second statistician's analysis is to be accepted.

Of course, there are many other possible models representing the underlying growth in the absence of the new diet. Under these alternatives, neither statistician's argument will be valid. It is possible that there is indeed a differential effect of diet (contrary to the first statistician's conclusion) but that the estimate proved by ANCOVA is biased. The moral of Lord's story is simply that in the absence of additional information on natural growth, there may be alternative models consistent with the data. Unless we can specify a correct model, statistical methods cannot be counted on to compensate adequately.

Covariate Measured With Error

So far we have been assuming that the covariate is measured without error. Suppose now that X represents a fallible variable subject to the usual psychometric assumptions (see Lord & Novick, 1968). The issue of reliability is very complex. Many definitions of reliability have been offered in an attempt to quantify the intuitive notion that part of an observed score is attributable to random fluctuation rather than to a stable characteristic. The effects of measurement error on statistical adjustments have been widely discussed, and many solutions to these problems have been proposed (e.g., Cochran, 1968; DeGracie & Fuller, 1972; Lord, 1960; Porter, 1967; Stroud, 1972).

The usual formulation of the measurement error problem assumes correct model specification in terms of an underlying true score T. That is,

$$E(Z|T, Q) = \mu + \beta T. \tag{41}$$

The observed score X is related to T by

$$E(X|T, Q) = T. \tag{42}$$

We define the reliability of X by

$$r = \rho_{TX \cdot Q}^2 = \frac{\sigma_{T \cdot Q}^2}{\sigma_{X \cdot Q}^2}. \tag{43}$$

Of course, in general, the conditional correla-

tions and variances may differ across groups, but to understand the main issues it suffices to consider the simple case when they are equal. In this case, it can be shown (Cochran, 1968) that the relationship between Z and X is of the form

$$E(Z|X, Q) = \mu' + \beta'X + \delta'Q, \quad (44)$$

where

$$\beta' = r\beta. \quad (45)$$

The effect of measurement error is to reduce the within-group regression coefficient (and hence the total adjustment) by a factor r. So, for example, if the reliability is .8, ANCOVA will remove 80% of the initial bias.

Based on Equation 45, methods have been proposed to estimate r and to "correct" the ANCOVA. But as noted above, this approach assumes that the model is correctly specified in terms of the true score. In general, however, there is no reason to expect that because a variable is free of measurement error it is a complete confounding factor. Instead of Equation 41 we might have

$$E(Z|T, Q) = \mu + \beta T + \delta Q \quad (46)$$

for some nonzero value of δ. In this case Equation 45 is still valid, and the reduction in bias using X remains r times the reduction in using X. So we have

$$\pi_X = (1 - r)\pi_T, \quad (47)$$

where π_X and π_T represent the proportions of bias remaining when X and T are the covariates.

This formula implies a rather curious possibility. Suppose that using T as the covariate results in an overadjustment ($\pi_T > 1$). Then by attenuating the relationship between T and Z, X may actually reduce the absolute magnitude of the remaining bias, by pulling π back toward 0. Of course, with finite samples, the use of X also implies lower precision of estimates.

With large samples, however, a variable with low reliability may be an excellent covariate. In fact, it may even be a complete confounding factor. This will occur when

$$\pi_T = \frac{1}{1 - r}, \quad (48)$$

which means $\delta' = 0$, but $\delta \neq 0$.

Although this result may seem counterintuitive, it is helpful in understanding what our definition of a complete covariate does and does not imply. It does not necessarily imply high precision with finite samples or a perfect relationship between covariate and outcome within groups. It is simply the condition under which unbiased estimation of α is possible, and implies that the model assumed by ANCOVA is correctly specified.

Note also that viewed in terms of model specification, the fallibility of the covariate is not in itself the problem. Having a perfectly reliable covariate will not guarantee correct adjustment, and having a fallible covariate does not necessarily result in bias. A fallible covariate may even be a complete confounding factor. As Overall and Woodward (1977) have emphasized, if assignment is explicitly on the basis of a fallible variable, the ANCOVA will be unbiased. A perfectly reliable variable on the other hand may not be complete. The important question is not one of measurement error, but whether Equation 15 holds in terms of whatever covariate is actually used.

Regression Artifacts

Another issue that has resulted in much confusion is the so-called *regression effect*. Suppose we wish to compare the performance of two groups on an outcome measure. The groups, however, are thought of as sampled from different populations. To adjust for differences in the groups, matching is commonly used. The problem is seen in the following terms:

> In order to get a matched group when the two populations have different mean values, we must take individuals who fall relatively high on one population and match them with individuals who fall relatively low in the other. Since the individuals in each group *will regress toward their own population mean*, the regression in the two groups will be different. Upon another test, our groups will no longer be matched. (Thorndike, 1942, p. 91)

So the mean difference on the outcome scores will differ, even if there is no difference between treatments offered to the groups. This difference is sometimes called a *regression artifact*.

This problem arises most commonly when a pretest measuring the same dimension as the outcome is used as the matching variable. The

regression artifact may then be attributed to imperfect test–retest reliability of the measuring instrument. As Thorndike (1942) noted long ago, however, the issue is much more general:

> The fallacies with which we are here concerned may arise whenever the measure or measures by means of which the groups were matched have less than a perfect correlation with the measure of the experimental variable which is being studied. A more limited example of this is found in the less than perfect correlation between a test and a subsequent retest with the same instrument. However, our argument is more general than this, and holds whenever groups are matched upon one measure or group of measures and then studied with regard to their performance on other measures which do not have a perfect correlation with the matching variable. Since this is universally true in the matched-groups experiment, the points to be raised here are of quite general application. (p. 85)

This problem is particularly confusing because it is not obvious what population mean the individuals in a group can be expected to regress toward. Suppose we know that one group is all black and the other all white. We might expect the blacks to regress toward the mean for black children. However, suppose we know they are black and living in Boston. A different mean may be relevant. This argument can be continued indefinitely. Viewed from this perspective, the concept of regression has a rather mystical quality. A retest score is being pulled by an irresistible force toward some predetermined norm from which it has deviated. Consider this convoluted argument of Campbell and Erlebacher (1970) attempting to circumvent this ambiguity:

> In situations such as this where control samples are chosen to have pretest scores equivalent to experimental samples, the question may be asked "Since the Head Start children are an extreme group, why don't they regress toward the overall population mean just as much as do the matched controls?" Comparable questions emerge when psychotherapy applicants are matched with a control sample chosen to have equally maladjusted test scores (e.g., Campbell and Stanley, 1973; 1977; pp. 11–21 and 45–50). Why are these controls expected to regress to the population mean while the therapy applicants are not? An initial answer is that person-to-person matching on individual scores involves the misleading exploitation of score instability phenomena to a much greater degree than do the complex of processes which produced the Head Start sample or the psychotherapy applicants. These groups turn out to be extreme when measured, but were not selected on the basis of their extreme scores. It is selection on the basis of extreme individual scores that creates most strongly the conditions under which obtained scores become biased estimates of true scores. (pp. 195–196)

From our standpoint, the reference to hypothetical populations is unnecessarily confusing. In any actual situation, the entire study sample may be viewed as drawn randomly from a single population to which generalizations are desired. Moreover, because we are ignoring finite-sampling issues, we can treat the sample and this population as identical.

Each individual in the population is seen as characterized by many variables. In particular, there is a value Y_i, Z_i, X_i, Q_i for individual i. In this formulation Z_i is unobservable for those individuals assigned to the treatment group, but is defined unambiguously.

With this formulation, the regression artifact is simply the bias resulting from statistical adjustment when X is not a complete confounding factor. Rather than saying that Z regresses toward two different population means, depending on which population the subject comes from, we can say that the subject's expected value of Z conditional on X depends on Q, that is,

$$E(Z|X, 1) \neq E(Z|X, 0). \tag{49}$$

An important special case that has led to much confusion occurs when Z represents a retest score on an instrument that is identical to that used for the initial test X. Suppose we are in a nonintervention situation, so that

$$Y = Z. \tag{50}$$

Then $\rho_{YX \cdot Q}$ can be interpreted as the test–retest reliability, r.

$$\rho_{YX \cdot Q} = \rho_{ZX \cdot Q} = r. \tag{51}$$

Now from Equation 28 we have

$$\pi = 1 - r. \tag{52}$$

This may be interpreted as an instance of standardized parallel growth. The underadjustment, or regression effect, is then seen as a natural consequence of this underlying growth model. On the other hand, we can assume that there exists an underlying true score T such that if the adjustment used T rather than X, there would be no remaining bias. The regression effect is then viewed as the result of measurement error.

Neither of these interpretations is *the* correct one. They are simply alternative ways of explaining the fact that the pretest does not contain complete information for adjusting posttest differences. What matters is not measurement error, differential growth, or regression toward a population mean, but simply that the relationship between posttest and pretest will not in general satisfy Equation 10.

Summary and Conclusion

I have tried in this article to provide a unified perspective on the problems in using statistical adjustments with uncontrolled selection. I have indicated how various problems can be viewed as special instances of model misspecification. In particular, problems raised by individual growth, measurement error, and regression effects can be understood in these terms.

For linear adjustment (ANCOVA), I derived an expression for π, the proportion of selection bias remaining after adjustment. This formula expresses π as a simple function of three intuitively meaningful parameters. One of these parameters, ρ_{ZQ}, expresses the relationship between the assignment process and the outcome that would be observed in the absence of the treatment. This correlation can have a substantial effect on π and cannot be estimated from data. So in general, we cannot know whether the remaining bias is likely to be small.

Unless we have evidence that X is a complete confounding factor, in the sense discussed above, we cannot be sure that all bias has been removed. As Lord (1967) has put it:

> With the data usually available for such studies, there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups. The researcher wants to know how the groups would have compared if there had been no preexisting uncontrolled differences. The usual research study of this type is attempting to answer a question that simply cannot be answered in any rigorous way on the basis of available data. (p. 305)

Many have interpreted this pessimistic conclusion as implying the need for more randomized experiments. Since adjustments cannot be counted on, methodologists must press for strict experimental control (see Campbell & Boruch, 1975).

In my view, this attitude is unrealistic. Particularly with large-scale social interventions, randomization is rarely practical. For most of our knowledge we must continue to rely on uncontrolled studies. So the crucial question is whether such studies can be better designed and analyzed. If, in Lord's words, "the data usually available for such studies" are inadequate, can we collect other data, not usually available, that will allow valid and useful inferences?

The answer depends in part on our willingness to modify the way we think about uncontrolled studies. The emphasis placed on randomized experimentation as an ideal mode of inquiry has led researchers to view alternative designs in terms of their closeness to this ideal, and to retain the idea of a treatment group versus control group comparison even when the groups have not been selected randomly. Statistical adjustments are employed in an effort to simulate the results that would have been observed under a randomized experiment. But as we have seen, the results of such attempts may be quite misleading.

Under randomization, the control group's performance is a proxy for that attained by the treatment group in the absence of intervention. With uncontrolled selection, the control group loses this property, and the rationale for such a group is greatly weakened. What is needed is a valid standard of comparison representing the outcomes for the treatment group had the treatment not been received. Information on the performance of a nonequivalent control group is relevant only if we are convinced that the groups do not really differ, or if we can specify a complete confounding factor X. Without randomization the burden of proof shifts to the investigator, who must provide evidence that two individuals with identical values of X, but assigned to different groups, would have the same outcome (on the average) in the absence of intervention.

On the other hand, it may be possible in some situations to estimate directly the performance of the treatment group in the absence of intervention, or more generally under alternative treatment conditions. This estimate can then serve as the standard of comparison against which to compare actual outcomes. For example, in some situations it

is possible to measure a characteristic of the treated population repeatedly before and after an intervention. Such repeated-measure, or time-series, designs (Campbell & Stanley, 1966; Glass, Willson, & Gottman, 1975) may allow strong causal inferences. The main threat to the validity of such designs is the possibility of a concurrent uncontrolled change at the time of the intervention. However, by applying techniques developed in single-subject research (Hersen & Barlow, 1976; Sidman, 1960), such as multiple-baseline and reversal designs, this problem can often be overcome.

Another promising approach, applicable when the outcome is a measure of developmental level, has recently been proposed (Bryk & Weisberg, 1976; Strenio, Bryk, & Weisberg, Note 3). *Value-added analysis* uses the variation in pretest scores for the treatment group to predict growth in the absence of intervention. This use of cross-sectional data to simulate the unobservable development that would have occurred without the treatment requires some strong assumptions, but these can in principle be tested using observable data. Extensions of the value-added idea to designs encompassing multiple measurements may ultimately allow valid estimates of effects on individual subjects (Strenio, Weisberg, & Bryk, Note 4).

Finally, historical information on the treated population as well as data on other relevant populations may be combined to yield reasonable standards of comparison. Bayesian and empirical Bayes approaches offer great promise as a formal technology for combining such sources of evidence (Rubin, 1978).

These approaches are not fully adequate as they now stand. They need to be refined to the point where they can produce standards of comparison that are valid under empirically testable assumptions. Attempting to develop such designs ought to be a top priority of evaluation methodologists. Until we have tried to develop alternatives not based on "approximations" to randomization, we should be cautious in discounting the value of uncontrolled studies. While statistical adjustments are certainly problematic, the potential contribution of uncontrolled studies has not really been tested.

Reference Notes

1. Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G. *Analysis of covariance in nonrandomized experiments: Parameters affecting bias.* Occasional Paper of the Stanford Evaluation Consortium, Stanford, Calif., 1977.
2. Goldberger, A. S. *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion Paper 123–72). Madison, Wis.: University of Wisconsin, Institute for Research on Poverty, 1972.
3. Strenio, J. F., Bryk, A. S., & Weisberg, H. I. *An individual growth model perspective for evaluating educational programs.* Paper presented at the meeting of the American Statistical Association, Chicago, August 1977.
4. Strenio, J. F., Weisberg, H. I., & Bryk, A. S. *Combining cross-sectional and longitudinal data to estimate individual growth curves.* Paper presented at the meeting of the American Statistical Association, San Diego, August 1978.

References

Anderson, S., et al. *Statistical methods in comparative studies.* New York: Wiley, in press.

Barnow, B. S., & Cain, G. G. A reanalysis of the effect of Head Start on cognitive development: Methodology and empirical findings. *Journal of Human Resources*, 1977, *12*, 177–197.

Bryk, A. S., & Weisberg, H. I. Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, 1976, *1*, 127–155.

Bryk, A. S. *An investigation of the effectiveness of alternative adjustment strategies in the analysis of quasi-experimental growth data.* Unpublished doctoral dissertation, Harvard University, 1977.

Bryk, A. S., & Weisberg, H. I. Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin*, 1977, *84*, 950–962.

Cain, G. G. Regression and selection models to improve non-experimental comparisons. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs.* New York: Academic Press, 1975.

Campbell, D. T. Reforms as experiments. *American Psychologist*, 1969, *24*, 409–429.

Campbell, D. T., & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs.* New York: Academic Press, 1975.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: A national debate. Vol. 3: Disadvantaged child.* New York: Brunner/Mazel, 1970.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.

Cochran, W. G. Errors of measurement in statistics. *Technometrics*, 1968, *10*, 637–666.

Cochran, W. G., & Cox, G. M. *Experimental designs* (2nd ed.). New York: Wiley, 1957.

Cochran, W. G., & Rubin, D. B. Controlling bias in observational studies: A review. *Sankhyā, The Indian Journal of Statistics*, Series A, 1973, *35*, 417–446.

Cohen, D. K. The value of social experiments. In A. M. Rivlin & P. M. Timpane (Eds.), *Planned variation in education: Should we give up or try harder?* Washington, D.C.: Brookings Institution, 1975.

Cox, D. R. *Planning of experiments*. New York: Wiley, 1958.

Cronbach, L. J., & Furby, L. How we should measure "change"—Or should we? *Psychological Bulletin*, 1970, *74*, 68–80.

DeGracie, J. S., & Fuller, W. A. Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. *Journal of the American Statistical Association*, 1972, *67*, 930–937.

Fisher, R. A. *Statistical methods for research workers* (4th ed.). Edinburgh, Scotland: Oliver & Boyd, 1932.

Gilbert, J. P., Light, R. J., & Mosteller, F. Assessing social innovation: An empirical base for social policy. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs*. New York: Academic Press, 1975.

Glass, G. V, Willson, V. L., & Gottman, J. M. *Design and analysis of time series experiments*. Boulder: Colorado Associated University Press, 1975.

Hersen, M., & Barlow, D. H. *Single-case experimental designs: Strategies for studying behavior change*. Elmsford, N.Y.: Pergamon Press, 1976.

Kempthorne, O. *Design and analysis of experiments*. New York: Wiley, 1952.

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin*, 1975, *82*, 345–362.

Lord, F. M. Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, 1960, *55*, 307–321.

Lord, F. M. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 1967, *68*, 304–305.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

McNemar, Q. *Psychological statistics* (4th ed.). New York: Wiley, 1969.

Overall, J. E., & Woodward, J. A. Nonradom assignment and the analysis of covariance. *Psychological Bulletin*, 1977, *84*, 588–594.

Porter, A. C. *The effects of using fallible variables in the analysis of covariance*. Unpublished doctoral dissertation, University of Wisconsin, 1967.

Riecken, H., & Boruch, R. F. (Eds.). *Social experimentation: A method for planning and evaluating social intervention*. New York: Academic Press, 1974.

Rubin, D. B. Matching to remove bias in observational studies. *Biometrics*, 1973, *29*, 159–183. (a)

Rubin, D. B. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 1973, *29*, 185–203. (b)

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974, *66*, 688–701.

Rubin, D. B. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 1977, *2*, 1–26.

Rubin, D. B. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 1978, *6*, 34–58.

Sidman, M. *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books, 1960.

Stroud, T. W. F. Comparing conditional means and variances in a regression model with measurement errors of known variances. *Journal of the American Statistical Association*, 1972, *67*, 407–412.

Suchman, E. A. *Evaluative research*. New York: Russell Sage Foundation, 1967.

Thorndike, R. L. Regression fallacies in the matched groups experiment. *Psychometrika*, 1942, *7*, 85–102.

Weiss, C. H. Evaluating educational and social action programs: A "treeful of owls." In C. H. Weiss (Ed.), *Evaluating action programs: Readings in social action and research*. Boston: Allyn & Bacon, 1972.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Received May 30, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 5, 1165–1168

Comparing Significance Levels of Independent Studies

Robert Rosenthal
Department of Psychology and Social Relations,
Harvard University

Donald B. Rubin
Department of Statistics,
Harvard University

Methods for comparing two or more statistical significance (p) levels are described; these methods are more rigorous, systematic, and informative than the comparisons that are commonly made by using a significant/not significant dichotomy. Formulas are provided for calculating the significance level of a comparison between two or more p levels.

Suppose that we wish to compare the results of two studies. If all that is reported for each is that the results are significant ($p \leq .05$) or not significant ($p > .05$), then the conclusion must be simply that the results are the same (both are significant, or neither is significant) or that the results differ (one is significant, and one is not significant). Most studies, however, do report information adequate for calculating p values (Rosenthal, 1978). The purpose of this article is to show how p values can be directly compared by calculating the significance level of the comparison.

Before describing our methods, two comments are in order. First, we could dichotomize each p value to be significant or not significant and use the crude method of comparison just outlined. This approach is clearly unwise because it does not use the more detailed information available in the p values. For example, a p of .05 tells about the same story as a p of .06 if both results are in the same direction and the studies are of similar size. This is true, even though many psychologists give far more credence to .05-level results than they do to .06-level results (Rosenthal & Gaito, 1963).

Our second comment on comparing ps is that if raw data or appropriate summary statistics are available, comparisons can be made that are more specific than those made solely from the p values. For example, raw effect sizes (e.g., mean differences), within-group variances, and residual variances can be directly compared with significance levels calculated for the comparisons. Since a p level is affected by raw effect size, residual variance, and sample size, the comparison of p levels is sensitive to differences in any of these components. In practice, particular areas of research tend to have reasonably homogeneous sizes of experiments, and there is substantial correlation between level of significance and magnitude of effect. For example, for eight research areas recently summarized (Rosenthal, 1976), the median correlation between effect size as measured by Cohen's d (Cohen, 1977) and the Z of the significance level was .74. Consequently, in many cases, comparisons of ps can be thought of as rough comparisons of effect sizes. Even when the sample sizes vary across the studies, however, the results presented here validly compare p values.

Notation

Consider first the simple case of two experiments, each with two treatments. Let p_j be the observed significance level in the jth experiment, and let $Z_j = Z(p_j)$ be the standard normal deviate corresponding to p_j. We assume

---

Preparation of this article was supported in part by a Fellowship from the John Simon Guggenheim Memorial Foundation to Donald B. Rubin, and by the Milton Fund of Harvard University, Biomedical Sciences Support Grant 5S07 RR 07046-12 from the National Institutes of Health to Harvard University, and the National Science Foundation.

The order of authors was determined alphabetically. Donald B. Rubin is on leave from the Educational Testing Service, Princeton, New Jersey.

Requests for reprints should be sent to Robert Rosenthal, Department of Psychology and Social Relations, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, Massachusetts 02138.

that the p_j and the Z_j are directed, so that if one study shows Treatment 1 superior and the other shows Treatment 2 superior, then one p_j will be less than .5 and the other greater than .5, and one Z_j will be positive and the other negative.

Suppose that Δ_j is the parameter to be estimated in the jth experiment: $\Delta_j = \mu_{1j} - \mu_{2j}$ where μ_{ij} is the population mean of Treatment i in Experiment j.

Let $\hat{\Delta}_j = \bar{Y}_{1j} - \bar{Y}_{2j}$ be the estimate of Δ_j in the jth experiment, where $\bar{Y}_{ij}$ is the sample mean of Treatment i in Experiment j. Throughout, the symbol $\hat{}$ will be used to indicate an estimate. The standard error of $\hat{\Delta}_j$ is $\sigma_{\hat{\Delta}_j}$; if the variances of the observations in Treatments 1 and 2 are the same, say σ_j^2, then $\sigma_{\hat{\Delta}_j}$ can be written as $\sigma_j\sqrt{(1/n_{1j}) + (1/n_{2j})}$, where n_{ij} is the number of observations comprising $\bar{Y}_{ij}$. The usual estimate of σ_j^2 is the residual mean square in the jth experiment.

Comparing Two Studies

We shall now give the results and examples of their applications. The technical discussion will be postponed until the end of the article.

Result 1

Suppose

$$\frac{\Delta_1}{\sigma_{\hat{\Delta}_1}} = \frac{\Delta_2}{\sigma_{\hat{\Delta}_2}};$$

that is, suppose that the quantities estimated by the t statistics of the two experiments are the same.

Then, for large n_{ij},

$$\frac{Z_1 - Z_2}{\sqrt{2}}$$

is distributed as a standard normal deviate.

Note that the test implied by Result 1 will be sensitive to different kinds of differences between the experiments. For example, if the precisions of the two studies are the same, that is, if $\sigma_{\hat{\Delta}_1} = \sigma_{\hat{\Delta}_2}$, then the test will be able to detect the difference in raw effects. For another example, if the raw effects are nonzero but equal, that is, if $\Delta_1 = \Delta_2 \neq 0$, then the test will be able to detect different precisions; the different precisions can be due to different residual variances σ_j^2 or different sample sizes.

Table 1
Example of Comparing Two p *Levels*

Study	One-tailed p	Z
A	$1/10^7$	5.20
B	.007	2.45
Difference		2.75*

Note. Sufficiently accurate ps can usually be obtained by interpolation or by using extended tables (e.g., Federighi, 1959).
* $2.75/\sqrt{2} = Z = 1.94$; $p = .026$, one-tailed.

Example 1. Table 1 shows the results of two studies of the effects of teachers' expectations on pupils' gains in intellectual performance (Rosenthal, 1976, p. 460). In both studies, gains in performance were greater when teachers had been led to expect better performance, but Study A showed results much more significant than those of Study B, the p levels being 10^{-7} and .007, respectively (one-tailed). It is of interest to compare these p values because the children of Study A were younger than the children of Study B. The comparisons were as follows:

From a normal table we find that $Z_A = 5.20$ and $Z_B = 2.45$. Thus $Z_A - Z_B = 2.75$ and $(Z_A - Z_B)/\sqrt{2} = Z = 1.94$, $p = .026$, one-tailed, suggesting more significant results for the younger children.

Example 2. Table 2 also shows the results of two studies of the effects of teachers' expectations on pupils' gains in intellectual performance (Rosenthal, 1976, p. 460; one of these studies has already been shown in Table 1). In this example, however, the results of one study were quite significant ($p = .007$), while those of the other were not ($p = .21$). Even though these p values are quite different, the difference between these p levels has a p value of only .123. One benefit of more sys-

Table 2
Example of Comparing Two p *Levels*

Study	One-tailed p	Z
A	.007	2.45
B	.21	.81
Difference		1.64*

* $1.64/\sqrt{2} = Z = 1.16$, $p = .123$, one-tailed.

Table 3
Example of Comparing Four p *Levels*

Study	One-tailed p	Z	Linear λ	Quadratic λ	Cubic λ
Grade level					
2	$1/10^7$	5.20	+3	+1	+1
3	.0001	3.72	+1	−1	−3
4	.21	.81	−1	−1	+3
5	.007	2.45	−3	+1	−1
Statistic					
Contrast Z			2.50	1.56	1.34
$Z^2 = \chi^2(1)$			6.23	2.43	1.79
One-tailed p			.006	.059	.090

tematic comparisons of p levels will be a decrease in the tendency to assume that two studies that differ in whether their results reach some conventional level of significance are really telling different stories about the state of nature.

Comparing Many Studies

Result 2 generalizes Result 1 to the case of many experiments. For Result 2, we suppose that there are K experiments and so let the index j run from 1 to K.

Result 2

Suppose

$$\frac{\Delta_1}{\sigma\hat{\Delta}_1} = \frac{\Delta_2}{\sigma\hat{\Delta}_2} = \cdots = \frac{\Delta_K}{\sigma\hat{\Delta}_K}.$$

Then for large samples,

$$\sum_{j=1}^{K} (Z_j - \bar{Z})^2$$

is distributed as χ^2 with $K - 1$ *df*, where $\bar{Z}$ is the mean of the Z_j.

Example 3. Table 3 shows the results of four studies of teacher expectations. The first and last studies listed were of second and fifth graders, respectively, and we met them in Table 1. We met the fourth graders in Table 2. The question we now ask is whether the four p levels of Table 3 are significantly different from one another. The sum of the squares of the deviations about the mean Z (*SS*) was computed to be 10.45, which is referred to the distribution of χ^2 with 3 *df*. For the four Zs of Table 3, we found p to be .015. Alternatively, the mean square (*SS/df*) is referred to the distribution of F with 3 and ∞ *df*. For the four Zs of Table 3, we found *SS/df* to be $3.48 = F(3, \infty)$, $p = .015$.

Contrasts in the Studies

Although we know how to answer the broad question of the significance of the differences among a collection of p levels, we may often be able to ask a more focused and more useful question. For the four studies of Table 3, for example, we are far more interested in the more focused question of whether lower p levels are found more at lower grade levels. Result 3 handles such questions.

Result 3

Suppose

$$\sum_{j=1}^{K} \lambda_j \frac{\Delta_j}{\sigma\hat{\Delta}_j} = 0,$$

where

$$\sum_{j=1}^{K} \lambda_j = 0.$$

Then for large samples,

$$\frac{\sum_{j=1}^{K} \lambda_j Z_j}{\sqrt{\sum_{j=1}^{K} \lambda_j^2}}$$

is distributed as a standard normal deviate.

Example 4. The column labeled Linear λ of Table 3 gives the weights of a linear contrast to address the question of whether lower p

values are found more often at lower grade levels. The analysis showed a clear linear trend for younger children to be more significantly affected by teacher expectations, $\Sigma(\lambda_j Z_j)/\sqrt{\Sigma\lambda^2} = 11.16/\sqrt{20} = Z = 2.50$, $p = .006$ (one-tailed).

We now know that the four p levels of Table 3 differ among themselves, $\chi^2(3) = 10.45$, $p = .015$, and that there is a linear trend for the p levels to be related to grade level, $Z = 2.50$, $Z^2 = \chi^2(1) = 6.23$, $p = .006$. Since the 3 df χ^2 of 10.45 is the sum of (a) the 1 df χ^2 of 6.23 corresponding to the linear trend and (b) an independent 2 df χ^2 corresponding to deviations from the linear trend, we have that $(10.45 - 6.23) = 4.22$ is distributed as χ^2 with 2 df.

In this case $\chi^2(2) = 4.22$, $p = .121$, suggesting no strongly significant curvilinear relationships. This χ^2 of 4.22 based on 2 df can be further split into a quadratic and a cubic component. The last two columns of Table 3 show the weights (orthogonal polynomials) employed for the quadratic and cubic contrasts and the Z, χ^2, and p for each. The $\chi^2(2)$ has been split into two $\chi^2(1)$s of 2.43 and 1.79, significant at $p = .06$ and .09, respectively. Since there is no theoretical reason to expect either a quadratic or a cubic trend for the present data, we might not normally split the 2 df χ^2 further, but there are applications where such further contrasts may be of value.

Technical Discussion

To prove Results 1, 2, and 3, fix $\delta_j = \Delta_j/\sigma\hat{\Delta}_j$ at some value, and let f_j, the df in each study, get larger and larger. As $f_j \rightarrow \infty$, $Z_j = Z(p_j)$ is the t test in the jth study and thus equals $\hat{\Delta}_j/\hat{\sigma}\hat{\Delta}_j$. Furthermore, in the limit, Z_j is normally distributed with mean δ_j and variance 1. Consequently, when $\delta_1 = \delta_2$, $(Z_1 - Z_2)/\sqrt{2}$ is $N(0, 1)$ (i.e., Result 1); and more generally, when $\Sigma\lambda_j\delta_j = 0$ (where $\Sigma\lambda_j = 0$), $\Sigma\lambda_j Z_j/\sqrt{\Sigma\lambda_j^2}$ is $N(0, 1)$ (i.e., Result 3). Also, when all $\delta_j = \Delta/\sigma\hat{\Delta}$, $\Sigma(Z_j - \bar{Z})^2$ is distributed as χ^2_{k-1}, because the Z_j are independent normal variables with common mean and variance 1 (i.e., Result 2).

To obtain insight into how large each f_j should be before we can trust the above asymptotic argument, we examine the mean and variance of Z_j for large but finite f_j. Using a Taylor series expansion of an expression in Wallace (1959, p. 1121), when f_j is large Z_j can be approximated as $t_j\,(1 - t_j^2/4f_j)$, where t_j is the t statistic that was used to look up the p value, p_j. Using this approximation and assuming normally distributed data, we can use Equation 1 in Johnson and Kotz (1970, p. 203), Stirling's approximation (Wilks, 1962, p. 177), and Taylor series expansions in $1/f_j$ to show that for large f_j, the mean of Z_j is approximately $\delta_j(1 - \delta_j^2/4f_j)$ and the variance of Z_j is approximately $[1 + (\frac{1}{2} - \delta_j^2)/f_j]$. These expressions suggest that if all f_j are large enough to ensure that all values of $\delta_j^2/4f_j$ and $(\frac{1}{2} - \delta_j^2)/f_j$ are close to zero (e.g., .10), then each Z_j will essentially have mean δ_j and variance 1, as with infinite f_j; as a consequence, we then expect the asymptotic arguments leading to Results 1, 2, and 3 to be appropriate. However, in rare circumstances when f_j is so small that $\delta_j^2/4f_j$ or $(\frac{1}{2} - \delta_j^2)/f_j$ is large (e.g., .5), then the significance tests presented in Results 1, 2, and 3 may be somewhat inexact, since even when all $\delta_j = \delta$, all the Z_j will not have (a) approximately the same mean or (b) variance equal to one.

References

Cohen, J. *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press, 1977.

Federighi, E. T. Extended tables of the percentage points of Student's *t*-distribution. *Journal of the American Statistical Association*, 1959, *54*, 683–688.

Johnson, N. L., & Kotz, S. *Continuous univariate distributions in statistics* (Vol. 2). Boston: Houghton-Mifflin, 1970.

Rosenthal, R. *Experimenter effects in behavioral research* (Enlarged ed.). New York: Irvington Press, 1976.

Rosenthal, R. Combining results of independent studies. *Psychological Bulletin*, 1978, *85*, 185–193.

Rosenthal, R., & Gaito, J. The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 1963, *55*, 33–38.

Wallace, D. L. Bounds on normal approximations to Student's and the chi-square distributions. *Annals of Mathematical Statistics*, 1959, *30*, 1121–1130.

Wilks, S. S. *Mathematical statistics*. New York: Wiley, 1962.

Received June 1, 1978

CONTENTS (continued)

Miller Appointed Editor, 1981–1986

The Publications and Communications Board of the American Psychological Association announces the appointment of George A. Miller as editor of *Psychological Bulletin* for the years 1981–1986. As of January 1, 1980, manuscripts should be directed to the Editor-elect:

George A. Miller
Department of Psychology
Princeton University
Princeton, New Jersey 08540

Ks Docket issue to the library 29/3/80

VOLUME 86 / NUMBER 6 NOVEMBER 1979

Psychological Bulletin

(Continued on inside back cover)

This issue completes Volume 86 and contains the author index to the volume.

PUBLISHED BIMONTHLY BY THE AMERICAN PSYCHOLOGICAL ASSOCIATION
ISSN 0033-2909

The *Psychological Bulletin* publishes evaluative reviews and interpretations of substantive and methodological issues in the psychological research literature. The Journal reports original research only when it illustrates some methodological problem or issue. Discussions of methodological issues should be aimed at the solution of some particular research problem on psychology, but should be of sufficient breadth to interest a wide readership among psychologists; articles of a more specialized nature can be directed to the various statistical, psychometric, and methodological journals. The *Bulletin* does not publish original theoretical articles; these should be submitted to the *Psychological Review*.

Abstracts: All articles must be preceded by an abstract of 100–175 words. Detailed instructions for preparation of abstracts appear in the *Publication Manual of the American Psychological Association* (2nd ed.), or they may be obtained from the Editor or from APA Central Office.

Blind review: Because reviewers have agreed to participate in a blind reviewing system, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet, which shows the title of the manuscript, the name of the author or authors, the author's institutional affiliation, and the date the manuscript is submitted. The first page of the manuscript should omit the author's name and affiliation but should include the title of the manuscript and the date it is submitted. Footnotes containing information pertaining to the author's identity or affiliation should be on separate pages. Every effort should be made to see that the manuscript itself contains no clues to the author's identity.

Manuscripts: Submit manuscripts in triplicate to the Editor, R. J. Herrnstein, *Psychological Bulletin*, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, according to instructions provided below.

Instructions to Authors: Authors should follow the directions given in the *Publication Manual of the American Psychological Association* (2nd ed.). Instructions on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors are requested to refer to the "Guidelines for Nonsexist Language in APA Journals" (Publication Manual Change Sheet 2, *American Psychologist*, June 1977, pp. 487–494) before submitting manuscripts to this journal. All manuscripts should be submitted in triplicate and all copies should be clear, readable, and on paper of good quality. Dittoed copies are not acceptable and will not be considered. Authors are cautioned to carefully check the typing of the final copy and to retain a copy of the manuscript to guard against loss in the mail.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1979: $40 domestic, $42 foreign, $7 single issue. APA member rate: $15. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Dept., APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in January) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036 and 1400 North Uhle Street, Arlington, Virginia 22201. Printed in the U.S.A. Second-class postage paid at Arlington, Va., and at additional mailing offices.

VOL. 86, NO. 6 NOVEMBER 1979

Psychological Bulletin

The Construct Validity of Egocentrism

Martin E. Ford
Institute of Child Development, University of Minnesota (Minneapolis)

Construct validation is briefly explained and then applied to egocentrism. Conceptual and operational referents of this construct are organized into three categories: visual/spatial egocentrism (what does the other see), affective egocentrism (what does the other feel), and cognitive/communicative egocentrism (what is the other thinking). Several kinds of reliability information are reported, and construct validity is evaluated primarily by examination of the relationships among measures of egocentrism within and between categories. Although interrater reliability and interrater agreement were found to be uniformly high for all egocentrism measures, and the measurement reliability was usually adequate, a few tasks were not internally consistent. Overall, the construct validity of egocentrism was not supported, since most task intercorrelations were low and often nonsignificant. From this evidence and an analysis of key egocentrism tasks, an alternative interpretation of the data based on cognitive constructs and task-specific and response-specific variables is proposed.

One activity essential for progress in theory and research is the validation of psychological constructs. Essentially, a construct is an unobservable characteristic of some entity, usually a person, that is hypothesized as an explanation for some observable phenomena. Constructs usually refer to an underlying psychological structure, state, system, or process, although these are sometimes not easily specified. Examples of some common psychological constructs that fit this description are intelligence, memory, anxiety, libido, identity, attachment, and, of course, egocentrism.

The author wishes to thank Andrew Collins and Daniel Keating for their help in editing the manuscript.

Requests for reprints should be sent to Martin Ford, Institute of Child Development, 51 East River Road, University of Minnesota, Minneapolis, Minnesota 55455.

Constructs that have been adequately validated through empirical testing can be efficient and reliable sources of guidance in our problem-solving activities. On the other hand, constructs that are not valid may distort our view of the relevant problems or may lead us to seek solutions in the wrong places. It is therefore important that the validity of constructs be evaluated. Cronbach and Meehl (1955) have attempted to explain when and why it is important to evaluate construct validity in the context of psychological testing. They comment that

> *construct validation* is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not "operationally defined" [i.e., there is no one operation that is by itself an adequate definition of the construct]. The problem faced by the investigator is, "What constructs account for variance in test performance?" (p. 282)

Cronbach and Meehl emphasize that construct validity can only be evaluated by

 0033-2909/79/8606-1169$00.75

integrating evidence from many different sources. They describe several kinds of investigations that can potentially provide information about the validity of a construct: (a) studies of group differences, (b) correlational and factor analytic studies, (c) studies of internal structure, (d) studies of change over occasions, and (e) studies of process.

In this article, studies that are relevant to the construct validity of egocentrism are reviewed. However, not all studies of egocentrism are included. Most of the research cited falls in the second and third categories described above, that is, correlational and internal structure studies. Where relevant to the discussion, studies of group differences (usually different age groups) are mentioned. However, since there are age differences on most measures with some cognitive content, these studies are given little weight. This kind of information is not as decisive or as discriminating as some other kinds of evidence might be and is therefore only selectively reported. For the same reason, studies of change over time might be less useful for answering questions about the construct validity of egocentrism, although this assertion is difficult to evaluate, since longitudinal studies in this area are so infrequent. Finally, studies of the psychological processes that explain or correlate with egocentric performance would probably be of great value, but unfortunately this kind of evidence is also sparse in the literature.

The article is organized around three issues:

1. What are the conceptual referents of the term *egocentrism?*

2. What are the various ways that researchers have attempted to operationalize this construct? Are these measures reliable?

3. Do the various operationalizations of egocentrism appear to be measuring the same underlying construct? If no, then what does each measure?

The present article concludes that the data do not support the construct validity of egocentrism and that several other sources of variance can be hypothesized to account for the data. Thus some additional evidence is presented that focuses primarily on the plausibility of this reinterpretation. The article relies, in particular, on an analysis of the requirements of several key egocentrism tasks and the kinds of errors that are commonly associated with them.

Conceptual Definitions of Egocentrism

Cronbach and Meehl (1955) assert that construct validation begins with a theory that defines the construct. They point out that if an investigator does not specify the meaning of the construct clearly enough, then others will be unable to evaluate the evidence provided for the validity of the construct. Although the nature of current psychological theorizing demands that a certain degree of vagueness be tolerated, the testability of a theory and the value of construct validation are greatly enhanced when this vagueness is minimized.

Another important point is that the term *egocentrism* is not always used as an explanatory construct. It is sometimes used simply to describe a specific social cognitive act and implies nothing about other acts or the underlying causes of the observed behavior. *Egocentric* may be useful as an adjective that describes a person's behavior in a given situation, but it should not be confused with the term *egocentrism* that is used to refer to a hypothetical underlying trait. It is only this latter use that is evaluated.

Looft (1972) attempted to pinpoint the meaning of egocentrism and concluded that it "does not pertain to selfishness or an overly keen regard of oneself, or even to the frequent use of 'I' or 'me.' The essential meaning of egocentrism is an embeddedness of one's own point of view" (p. 74). In Piaget's (1926) theory, which is the context in which the construct is typically used, egocentrism is defined as a lack of differentiation in some aspect of subject–object interaction. Feffer (1959, 1970) defines it as the inablity to "decenter," where decentration refers to one's ability to shift attention to consider more than one aspect of an event. Although each of these definitions is a little different, they share a common core of meaning: Each refers to an individual's failure to perceive a situation or an event in more than

one way. This one way of perceiving is the one that is easiest for the individual, that is, the one that requires no conceptual elaboration beyond what is directly perceived.

Conceptually and methodologically, the referents for egocentrism can be organized into three categories. Adapted from Shantz's (1975) review of social cognitive development, these are (a) What does the other see? (b) What does the other feel? and (c) What is the other thinking? Each of these incorporates an appreciation for perspectives other than one's own and the ability to infer in a given situation what these alternative perspectives are. The differences among these three categories are essentially in the domain each pertains to: visual/spatial (sometimes called perceptual), affective, and cognitive/communicative (sometimes called conceptual).

Piaget's theory and the current developmental literature imply that egocentrism is a characteristic of individuals that is both consistent across situations and stable over time. Because egocentric performance is typically regarded as an indication of the developmental status of the individual in terms of Piaget's stages of cognitive development, egocentrism is clearly meant to refer to a generalized trait. In other words, the presence or absence of this hypothetical construct is considered to be a sufficient explanation for the presence or absence of a wide range of phenotypically diverse behavior. The major prediction that one can make from this conceptualization is that if egocentrism is a unitary trait that can be used to predict failures in all kinds of perspective taking, then measures of egocentrism from each of the three domains previously described should be positively and significantly correlated. It is reasonable to expect that correlations of measures within a given domain would be higher than those between categories (due to domain specific variance), but both sets of correlations should exceed correlations between egocentrism measures and measures of other constructs. Although these latter correlations may be expected to be positive and significant for theoretically related constructs such as intelligence, conservation, and popularity (Rubin, 1973), they should not equal or exceed within-construct correlations.

In the section on Relationships Among Egocentrism Measures, studies that have computed these kinds of correlations are reviewed. First, however, the next section describes operational definitions (measures) of egocentrism and reports evidence on the stability and internal consistency of these measures.

Operational Definitions of Egocentrism

Measures of Visual/Spatial Egocentrism: What Does the Other See?

Piaget and Inhelder (1956) developed the first measure that was intended to assess the ability to imagine how an object or set of objects would appear from a different occupied position (i.e., from the perspective of another person). In their test, known as the three-mountains problem, a child sits in one of four chairs positioned around a table on which three mountain-shaped objects are placed. A doll is placed in one of the three vacant chairs, and the child indicates what the doll sees by pointing to one of a set of drawings or photographs, by drawing what the doll sees, or by recreating the doll's view by manipulating materials provided to the child. A correct response is taken as evidence for visual/spatial perspective-taking ability.

Many innovations on this task have appeared in the literature; various stimulus displays have differed fundamentally on dimensions of complexity and familiarity. Flavell, Botkin, Fry, Wright, and Jarvis (1968) developed several measures, including one in which four displays are presented one at a time in a standard sequence. Each display consists of a set of objects fastened to a board, which is placed in the middle of a small rectangular table. These displays are increasingly complex: The first has only a single red wedge of wood; the second consists of three vertically oriented blue cylinders of equal height; and the last two displays, one blue and the other half red and half blue, consist of three cylinders of unequal height. The child's task is to reconstruct the experimenter's visual perspective using a duplicate set of unfastened objects.

Many other visual/spatial role-taking tasks could be described; brief characterizations of some of these should suffice. In a simple task, Liben (1978) had young children wearing yellow sunglasses indicate how a white card would appear to an experimenter wearing green sunglasses. Coie, Costanzo, and Farnill (1973) used a display in which a toy doll was oriented in different positions with respect to three toy houses varying in both size and color. Kurdek and Rodgon (1975) created a display with presumably attractive and familiar Walt Disney characters. Photographs, including several on a photo cube, and a table setting were used by Zahn-Waxler, Radke-Yarrow, and Brady-Smith (1977), based on measures developed by Flavell et al. (1968). Fishbein, Lewis, and Keiffer (1972) displayed either one or three toys to a child, who either pointed to one of a set of four or eight photographs or actually turned the display until the correct view was facing towards the child. Eliot and Dayton (1976) used an adaptation of the three-mountains task; by varying the shape of the stimulus objects (blocks) and the shape of the board that supported them and by varying as well the arrangement of the objects, 39 different configurations could be constructed. Finally, Shantz and Watson (1971) had young children view a display in a covered box and then view the same display from the opposite side; on some trials the display was rotated 180° and the children's expressions of surprise or amusement were noted.

It is important to note that in all of these measures, only certain responses can be interpreted as a manifestation of egocentrism, that is, those that represent the actual visual/spatial perspective of the subject. All other errors are nonegocentric errors because they represent perspectives other than those experienced by the subject. For example, it would not be evidence of egocentrism if a child, when asked to show how someone sitting opposite the child would see a rectangular display, were to indicate a perspective that was incorrect but different from his or her own, such as the perspective of someone sitting to the immediate right. No matter how poorly the child performs on the task, egocentrism can only logically be inferred when the child's own perspective is offered as the correct answer. This point is elaborated in a later section in which age differences in visual/spatial perspective-taking errors are considered.

Reliability. Only one estimate of the reliability of any measure of visual/spatial egocentrism appears in the published literature. Rubin (1973) reports that test-retest (TR) reliability for the Flavell et al. (1968) measure is between .85 and .95 for his sample of 5- to 12-year-olds. Where reported, the interrater reliability and interrater agreement have also been in this range (Rubin, 1973; Zahn-Waxler et al., 1977).

Measures of Affective Egocentrism: What Does the Other Feel?

Affective egocentrism refers here to an inability to infer the feelings of others. It does not imply the ability or inclination to share these feelings, which distinguishes it from a related construct, empathy. (In general, taking the perspective of another person requires being able to identify and appreciate a different view of the world in some domain. It does not require that this perspective be experienced in precisely the same way; indeed, that would be unlikely, since immediately perceived and inferred experience would have to be equivalent.)

In one widely used measure of affective egocentrism, Borke's (1971, 1973) Interpersonal Perception Test, a child is told a short story that portrays an emotionally stimulating situation, such as losing a pet (sadness), going to a birthday party (happiness), having a toy broken by another child (anger), and being lost (fear). There are 23 stories in all. Presented along with these stories is a picture of the described situation, in which the appropriate character has a blank face. The child's task is to supply the proper facial expression. A similar but even simpler measure was used by Feshbach and Roe (1968), whose Affective Situations Test provided the appropriate facial expressions for the story characters along with the situational cues. An important criticism of these

measures is that one cannot be sure that correct responses are evidence of true perspective taking, since there is no clear criterion for discriminating between subjects' attributing their own responses to a situation and actually inferring the emotional responses of another.

A measure of affective egocentrism that may permit this discrimination was developed by Rothenberg (1970). This task differs from Borke's Interpersonal Perception Test on dimensions of familiarity and to a lesser degree, complexity. For this measure, children are required to judge the feelings of individuals unlike themselves (adults) in relatively unfamiliar situations (e.g., unexpectedly bringing friends home for dinner to an unprepared spouse). This task is more complex in that no visual cues are available and the taped verbalizations are more adultlike and presumably more difficult to comprehend.

Several researchers (Burns & Cavey, 1957; Deutsch, 1974; Kurdek & Rodgon, 1975) have used the general strategy of presenting pictures or films of situations in which the facial expression of the central character is not what one would expect on the basis of the contextual cues (e.g., frowning at a birthday party or following a helpful gesture by another). Inferring affective perspective taking in these tasks is difficult, since children with the ability to infer the feelings of others may give differing responses depending on individual differences in the salience of the incongruent facial and situational cues. The problem mentioned earlier of discriminating attribution of one's own feelings and inferring others' feelings is here also.

Reliability. Only three studies report measurement reliabilities for affective egocentrism, and all are dangerously low. For a sample of third and fifth graders, the internal consistency of the Rothenberg measure has been reported as .28–.47 (Rothenberg, 1970), .30 (Hudson, 1978), and .50 (Rubin, 1978). In the latter study, the internal consistencies within Grade Levels 1, 3, and 5 were only .18, .20, and .39, respectively. Again, where reported, interrater reliability and interrater agreement are fairly high (Hudson, 1978; Moir, 1974; Rothenberg, 1970; Rubin, 1978).

Measures of Cognitive/Communicative Egocentrism: What Is the Other Thinking?

This is the broadest of the three categories and is therefore not as conceptually distinct as the previous two. In general, this set of tasks requires the subject to infer something about the thoughts, motives, or intentions of another person. Sometimes this is done more or less in the absence of the subject's potentially interfering cognitions, but more commonly the evidence of interest is whether one can overcome the tendency to attribute one's own knowledge to another in some sort of communicative situation.

Referential communication. Perhaps the most widely used measures of cognitive/communicative egocentrism are those used in referential communication studies, in which a listener must select the appropriate object (referent) from a set of objects (nonreferents) on the basis of a verbal message by a speaker. In some cases the major dependent variable is the number of successful matches made by the speaker and listener, although since the characteristics of the listener can vary considerably, others have been more interested in the actual content of the speaker's message. This also provides a more direct assessment of the subject's communicative ability. The most widely used measure of referential communication ability was devised by Glucksberg and his colleagues (Glucksberg & Krauss, 1967; Glucksberg, Krauss, & Higgins, 1975; Glucksberg, Krauss, & Weisberg, 1966). In this task, a speaker and a listener are seated at opposite ends of a table with an opaque screen separating them. Several stimuli are pictured on cards or blocks; these stimuli are novel graphic designs that are difficult to label or describe. The speaker then attempts to provide descriptions of each stimulus so that the listener can successfully discriminate the referent from the nonreferents. Successive trials are not independent in the sense that the set of nonreferents diminishes in size with each trial, although some researchers have used the full set of stimuli for each trial. A variant on the general Glucksberg et al. procedure is to have listeners provide feedback indicating that they didn't understand the first message and then

to assess how the speaker's communication is affected. This manipulation is based on the premise that egocentric speakers will be unable to recode the message to take into account the communicative failure.

Other measures have been devised to assess referential communication ability, and in a situation analogous to that for the visual/spatial and affective egocentrism measures, these differ essentially in the familiarity and perceptual complexity of the stimulus items used. In addition to the Krauss and Glucksberg Blocks Task (Krauss & Glucksberg, 1969), Piché, Michlin, Rubin, and Johnson (1975) administered the Baldwin and Garvey Picture Identification Task (Baldwin & Garvey, Note 1), in which subjects describe a picture of a Dr. Seuss-like animal to a listener who must select the correct picture from a set of seven that closely resemble each other. Piché et al. (1975) also administered the Crystal Climbers Task, in which subjects describe a model made of white plastic circles, squares, rectangles, and cylinders of various sizes to a listener who must construct the same model out of a set of unassembled pieces. Other studies used measures of less perceptual and conceptual complexity. Shatz and Gelman (1973) had their subjects describe which of several airplanes of similar appearance to choose. Maratsos (1973) had young children describe to an experimenter who either was watching or who could apparently not see (i.e., hands over eyes but with a small crack to peek through) which of several toys to place in a toy car that the child was to catch as it rolled down a small hill. These toys consisted of familiar, easily discriminable and describable objects: a red duck, a green duck, small dogs, and boys and girls. Hoy (1975) also varied whether the listener could see the speaker, but in addition the familiarity and complexity of the objects to be described were manipulated (i.e., a horse or a random shape).

Social and private speech. Another method of assessing cognitive/communicative egocentrism is to observe the quality of children's speech in naturalistic settings. These settings may be social, in which case the data of interest are the degree to which speakers can modify their messages according to the status of their listeners, especially those who are younger and presumably less competent communicators (Garvey & Hogan, 1973; Shatz & Gelman, 1973). In other instances, these settings may be mostly nonsocial in the sense that any speech that occurs is not directed at another person. This measure clearly originated from Piaget's (1926) characterization of the speech of young children as repetitious and like a "collective monologue." Kohlberg, Yaeger, and Hjertholm (1968) developed a quantitative scale in which certain kinds of private speech, such as the repetition of words for their own sake or simple descriptions of one's own ongoing activity, were rated as more egocentric than other kinds of private speech, such as speech that was used to guide and control activity or inaudible muttering. This kind of measure seems to be an indirect means of assessing the ability to infer the cognitive perspective of another and to adjust one's behavior accordingly. Social speech measures might be better suited to this purpose. However, Kohlberg et al. comment that self-communication may not be very different from social communication with someone you know intimately.

Feffer's Role-Taking Task. A third popular kind of cognitive/communicative egocentrism measure is Feffer's Role-Taking Task (RTT; Feffer, 1959, 1970; Feffer & Gourevitch, 1960; Feffer & Suchotliff, 1966). This measure, which attempts to tap an individual's ability to decenter or to see an interpersonal situation from the perspective of another, requires subjects to make up an initial story as one character would tell it from a picture (e.g., a Thematic Apperception Test card) that portrays at least three characters. Then they must retell the story as it would be perceived or experienced by the other characters, repeating the story once for each character. Scoring is based on the extent to which the sequential story telling reflects an ability to logically coordinate the different versions of the initial story and to elaborate the roles and internal states of the characters (Schnall & Feffer, Note 2).

Privileged information. A fourth measurement strategy in this domain is the privileged

information paradigm originated by Flavell et al. (1968) and further developed by Chandler (1973). Flavell's version, known as the apple-dog story, is similar to Chandler's several stories. Subjects first describe a story that is displayed in a sequence of several cartoons that depict story characters in emotionally charged and stressful situations. They then must retell the story from the perspective of a bystander who has no knowledge of the activity portrayed in an important subset of the cartoons. This knowledge is crucial for understanding the outcome of the story and is the subject's "privileged information" that must not be attributed to the bystander if successful role taking is to occur. Scoring is based on the degree to which subjects are able to restrict their second narrative to the limited perspective of the bystander. Ambron and Irwin (Note 3) used a similar strategy, except that in their test the narrative is supplied by the experimenter, and there are fewer cartoons (four).

Chandler, Helm, and Smith (Note 4) developed a task based on "droodles," which are clever drawings that convey some situation with a minimum of lines. Subjects first saw only a part of the droodle, which by itself was uninterpretable, and then were given the whole drawing. The task was then to interpret the picture from the perspective of someone who had only seen the uninterpretable part of the droodle.

Zahn-Waxler et al. (1977) used a wide array of tasks that assessed conceptual role taking, including two privileged information stories. However, they also included tasks that involved (a) choosing an appropriate birthday gift for other persons; (b) choosing the appropriately sized chair for one's self and for an adult experimenter; (c) indicating which of two games, including one that a confederate preferred and the subject did not prefer, that the confederate would like to play with; and (d) indicating which of two foods (attractive cookies or juice with soggy crackers) that a confederate who pretended to hate cookies would prefer to eat.

Recursive thought. Another means used to assess cognitive/communicative egocentrism is through the child's understanding of the recursive nature of thought (Miller, Kessel, & Flavell, 1970). This measure involves showing subjects several cartoonlike drawings in which scalloped cartoon clouds represent thinking and smooth cartoon clouds represent talking. The faces of a boy, a girl, a mother, and/or a father are pictured in the drawings. Different configurations of smooth and scalloped clouds are embedded in one large scalloped cloud, indicating that the depicted character is thinking about something. After subjects are trained to understand the meaning of the clouds, they are asked to describe what the person in the picture is thinking. The pictures to be described range from relatively simple situations (e.g., the boy is thinking about the girl) to fairly complex representations (e.g., the boy is thinking that the girl is thinking of him talking to her). Presumably, an inability to take the perspective of another will be reflected in an inability to describe events such as thinking about another's thoughts.

Infer game strategies. A sixth and final set of measures in this domain are those that require subjects to infer the strategy of an opponent in a game. The most widely used of these is the Flavell et al. (1968) nickel–dime game, in which the subject has a nickel and a dime and two upside-down cups with another nickel and dime taped to them. The subjects' instructions are to cover the coins with the pair of cups in such a way that their opponents will choose the lesser of the two rewards. For each trial, subjects may be asked to explain their rationale for the placement of the cups. These explanations are analyzed for the degree to which they reflect consideration of the thought processes (i.e., strategies) of the opponent. The hide-the-penny guessing game used by Selman (1971a, 1971b) is similar in purpose and execution.

Reliability. Relatively abundant reliability information is available for the various measures of cognitive/communicative egocentrism. The three studies that provide measurement reliabilities for referential communication tasks report TR correlations of .86 (Chandler, Greenspan, & Barenbaum, 1974), .89 (Deutsch, 1974), and .85–.95 (Rubin, 1973), which indicates that this is a

fairly reliable measure. Rubin also found that the reliability (TR) of the private speech and recursive thought measures was adequate (.85–.95). However, Kohlberg et al. (1968) reported a TR correlation of only .43 for their measure of private speech. Similarly low reliabilities are reported for Feffer's RTT. All four studies that reported reliability coefficients obtained low internal consistency estimates: .27 (Keller, 1976), .40 (Kurdek, 1977), .40 (Turnure, 1975), and .42 (Feffer & Gourevitch, 1960). Kurdek also obtained a TR correlation for this task of .60. On the other hand, Chandler's privileged information stories are more reliable, although not dramatically so. Internal consistency reliabilities of .91 (Chandler et al., 1974), .65-.86 (O'Connor, Note 5), .56 (Kurdek, 1977), and .52 (Rubin, 1978) have been reported. Chandler et al. and Kurdek obtained TR correlations of .84 and .68, respectively. And finally, O'Connor and Kurdek reported adequate internal consistency (.77–.85 and .68) for two different game strategy inference measures.

Interrater reliability and interrater agreement are uniformly high for all of the above measures, indicating that scoring is not a major problem (Byrne, 1974; Chandler & Greenspan, 1972; Chandler et al., 1974; Deutsch, 1974; Feffer, 1959; Hudson, 1978; Kohlberg et al., 1968; Kurdek, 1977; Leahy & Huard, 1976; Piché et al., 1975; Rubin, 1972, 1974, 1978; Turnure, 1975; Urberg & Docherty, 1976; Weinheimer, 1972; Wolfe, 1963; Zahn-Waxler et al., 1977; Marsh & Serafica, Note 6; Olejnik, Note 7). Unfortunately, some of these reliability estimates are difficult to evaluate, since the actual degree of rater independence is often unknown because the scoring procedures are not described in sufficient detail.

Summary. The various measures of egocentrism can be organized into three categories that correspond to their conceptual referents. The major dimensions that differentiate measures within categories are the familiarity of the task stimuli and the complexity of the task in terms of its perceptual characteristics, the task instructions, and the type of response required. The main dimension that differentiates measures between categories is the kind of inference required. (i.e., Is it concerned with seeing, feeling, or thinking?) Visual/spatial egocentrism measures are patterned after Piaget and Inhelder's (1956) three-mountains problem, but there are many variations ranging from simple displays with familiar objects to fairly complex displays with novel objects. Not all errors on these tasks are egocentric errors. Affective egocentrism measures are fewer in number, but these have special problems that relate to the kind of inferences that can legitimately be made about subjects' thought processes (e.g., memory vs. perspective taking). Also, the internal consistency of these measures appears to be lower than one would like, indicating that the shared variance among task items is minimal. There are many different kinds of cognitive/communicative egocentrism measures, including referential communication measures, observational measures of children's naturalistic speech, Feffer's RTT, measures that assess appreciation of privileged information and comprehension of recursive thought, and measures that assess awareness of another's strategies in a game situation. Where reported, reliabilities seem to be at least moderately high for all of the measures in this category except Feffer's RTT, the internal consistency of which is too low to claim that it reliably measures a unitary dimension.

Relationships Among Egocentrism Measures

Tables 1, 2, and 3 summarize the data reviewed in this section and the last section that is relevant to the construct validity and reliability of measures of egocentrism. In these matrices of correlations, four kinds of reliabilities are indicated. Two refer to the reliability of the measure itself, TR and internal consistency (IC). The other two refer to the reliability of the judges who scored the egocentrism measures, interrater reliability (IR) and interrater agreement (IA).

The row (and column) labeled Other conceptual role-taking tasks (Tables 1 and 3) refers to a wide variety of tasks not subsumed by the other categories. These include a word association test (Piché et al., 1975), one of

the many Flavell et al. (1968) tasks (Moir, 1974), the conceptual role-taking tasks described in the previous section that were used by Zahn-Waxler et al. (1977), and a task devised by Selman in which one must infer the thoughts and intentions of a filmstrip character (Kurdek, 1977). Included in Tables 1-3 are several constructs other than egocentrism, so discriminant validity can be evaluated as well as convergent validity.

The data presented in Tables 1–3 should be interpreted with great caution. These tables are an organizational device and not the final word on the construct validity of egocentrism. Correlations are not the only kind of relevant evidence. The data in Tables 1–3 will be more meaningful if one has a grasp of the rationale for each study, the specific procedures used, the situation in which the testing took place, and so forth, which can be obtained from the original reference. Nevertheless, these tables should be a convenient starting point for those interested in investigating the construct validity of egocentrism.

Visual/Spatial Egocentrism and Affective Egocentrism

Using two of the Flavell et al. (1968) visual/spatial measures and Rothenberg's affective egocentrism measure, Moir (1974) found in a sample of 40 11-year-old New Zealand girls essentially zero correlations between the two domains after IQ was partialed out. One of the two correlations was significant ($r = .36$, $p < .05$) before partialing out IQ. Rubin and Maioni (1975), using an adapted version of the three-mountains problem and Borke's affective egocentrism task, obtained a correlation of .44, which was not significant for their small sample of 16 preschool children. The only other study to relate measures in these two domains (Kurdek & Rodgon, 1975) correlated scores on a visual/spatial task that involved three Walt Disney characters and a task that required identification of affective states in situations in which facial and contextual cues were incongruent. In their large sample ($N = 167$) of children from Grades K, 2, 4, and 6, the correlation between the two tasks was significant only for second graders, and even there it did not account for a large portion of the variance ($r = .36$).

One can tentatively conclude that the proportion of variance shared by these two domains is small indeed and is possibly attributable to general intelligence. One certainly cannot conclude from these data that an underlying construct of egocentrism can explain individual differences on these tasks. This conclusion must remain tentative, however, because the internal consistency of the affective egocentrism measures is so low that it may preclude obtaining high validity correlations. Although this may be a function of the hypothesized trait being measured, it may be a measurement failure. The question of what visual/spatial and affective egocentrism tasks might be measuring if not egocentrism is dealt with later.

Affective Egocentrism and Cognitive/Communicative Egocentrism

Only a few studies have correlated measures in these two domains, and these substantially support the conclusions advanced previously. Using a sample of 73 delinquents and nondelinquent controls, Rotenberg (1974) found a correlation of .02 between a cognitive role-taking measure that assessed subjects' ability to predict other's everyday behavior and an affective role-taking measure that assessed their tendency to relieve the distress of others. Hudson (1978) reported a small but significant relationship ($r = .16$) between Rothenberg's affective egocentrism measure and Flavell's apple-dog story. Similarly, Rubin (1978) found partial correlations (controlling for chronological age) of $-.23$ and .00 between Chandler's privileged information stories and the Rothenberg and Borke measures, respectively. Rubin also reported mostly nonsignificant partial correlations between these affective egocentrism measures and the hide-the-penny game, the Glucksberg-Krauss task, and the Miller et al. recursive thought measure, although these latter two were significantly related to Rothenberg's task ($r = .25$ and .20, respectively).

Table 1
Construct Validity and Reliability of Measures of Egocentrism

Measure	Visual/spatial egocentrism	Affective egocentrism	Referential communication	Recursive thought
Visual/spatial egocentrism	TR = .85–.95 [*N* = 10, 5–12(24)] IR = .82–.95 [*N* = 20, 5–12(24)] IA = 86%–100% [*N* = 108, 3–7(36)]			
Affective egocentrism	.44 [*N* = 16, 3–5(27)] .36* (.01) [*N* = 40, 11(17)] .24 (.02) [*N* = 40, 11(17)] .36* at one grade, *ns* at other three grades [*N* = 167, 5–12(14)]	IC = .28–.47 [*N* = 40, 11(17)] IC = .30 [*N* = 110, 7–8(10)] IC = .50 [*N* = 12, 3–11(26)] IR = .91 [*N* = 108, 8–11(22)] IA = 90% [*N* = 110, 7–8(10)] IA = 85% [*N* = 40, 11(17)] IA = 100% [*N* = 12, 3–11(26)] IA = 81% [*N* = 12, 3–11(26)]		
Referential communication	.65* (.35*) [*N* = 80, 5–12(24)] .49* [*N* = 112, 8–76(25)] *sig* at one of two ages [*N* = 74; 8, 12(9)]	(.06) [*N* = 63, 3–11(26)] (.25*) [*N* = 142, 3–11(26)]	TR = .91 [*N* = 10, 5–12(24)] TR = .86 [*N* = 125, 8–15(4)] IR = .98 [*N* = 80, 5–12(23)] IR = .96 [*N* = 68, 9–12(15)]	
Recursive thought	.73* (.36*) [*N* = 80, 5–12(24)]	(.10) [*N* = 40, 3–11(26)] (.20*) [*N* = 46, 3–11(26)]	.72* (.31*) [*N* = 80, 5–12(24)] (.46*) [*N* = 116, 3–11(26)]	TR = .85–.95 [*N* = 10, 5–12(24)] IR = .82–.95 [*N* = 20, 5–12(24)] IC = 100% [*N* = 12, 3–11, (26)]
Feffer's Role-Taking Task	.35* [*N* = 30, 11(31)] .25 [*N* = 30, 7(31)] .00 [*N* = 30, 9(31)] *sig* at one of two ages [*N* = 74; 8, 12(9)]		.23 [*N* = 20, 9–10(20)] −.08 [*N* = 20, 9–10(20)]	
Privileged information	.52* at one grade, *ns* at other three grades [*N* = 167, 5–12(14)]	.16* [*N* = 110, 7–8(10)] (.00) [*N* = 38, 3–11(26)] (−.23*) [*N* = 38, 3–11(26)]	.44* [*N* = 20, 9–10(20)] .32 [*N* = 20, 9–10(20)] .31* [*N* = 125, 8–15(4)] .05 [*N* = 68, 9–12(15)] (−.25*) [*N* = 114, 3–11(26)]	(−.18*) [*N* = 114, 3–11(26)]
Infer game strategies	.58* (.43*) [*N* = 40, 11(17)] .31 (.15) [*N* = 40, 11(17)] *ns* [*N* = 60, 4–6(29)]	.49* (.35*) [*N* = 40, 11(17)] (.07) [*N* = 63, 3–11(26)] (.06) [*N* = 142, 3–11(26)]	(.05) [*N* = 142, 3–11(26)]	(.22*) [*N* = 114, 3–11(26)]
Private speech	.28* (−.06) [*N* = 80, 5–12(24)]		.37* (.07) [*N* = 80, 5–12(24)]	.32* (−.02) [*N* = 80, 5–12(24)]
Social speech	*sig* [*N* = 108, 507(14)]			
Other conceptual role-taking tasks	.41* (.27) [*N* = 40, 11(17)] .34* (.23) [*N* = 40, 11(17)] .34* [*N* = 54, 5–7(36)] .08 [*N* = 54, 5–7(36)]	.25 (.10) [*N* = 40, 11(17)] .02 [*N* = 73, 11–15(21)]	.01 [*N* = 20, 9–10(20)] −.03 [*N* = 20, 9–10(20)]	
IQ or mental age	.75* [*N* = 80, 5–12(24)] .25–.44* [*N* = 90, 7–11(31)]	.28* [*N* = 110, 7–8(10)]	.76* [*N* = 80, 5–12(24)] .43* [*N* = 55, 7(28)]	.77* [*N* = 80, 5–12(24)]

Table 1 (*continued*)

Measure	Visual/spatial egocentrism	Affective egocentrism	Referential communication	Recursive thought
Conservation	.63* (.26*) [N = 80, 5–12(24)] *sig* [N = 108, 5–7(1)]		.65* (.31*) [N = 80, 5–12(24)]	.65* (.26*) [N = 80, 5–12(24)]
Popularity	.68* [N = 16, 3–5(27)] .05 [N = 80, 5–12(24)] *sig* [N = 108, 5–7(1)]	.68* [N = 16, 3–5(27)]	.52* [N = 60, 3–5(5)] .22 [N = 60, 3–5(5)] .05 [N = 80, 5–12(24)]	−.11 [N = 80, 5–12(24)]

Note. TR = test–retest reliability; IC = internal consistency; IR = interrater reliability; IA = interrater agreement. Each entry is in the general form: r(partial r) [N = X, ages(reference key number)], where r is the correlation between two sets of scores, partial r is the same correlation with IQ, mental age, or chronological age partialed out, N is the number of subjects in the study, ages is the ages (in years) of groups used to compute the correlations, and reference represents a number in the following reference key: 1 = Bunting (1975); 2 = Byrne (1974); 3 = Chandler & Greenspan (1972); 4 = Chandler, Greenspan, & Barenbaum (1974); 5 = Deutsch (1974); 6 = Feffer (1959); 7 = Feffer & Gourevitch (1960); 8 = Feffer & Suchotliff (1966); 9 = Heilbrunn (1974); 10 = Hudson (1978); 11 = Keller (1976); 12 = Kohlberg, Yaeger, & Hjertholm (1968); 13 = Kurdek (1977); 14 = Kurdek & Rodgon (1975); 15 = Leahy &Huard (1976); 16 = Marsh & Serafica (Note 6); 17 = Moir (1974); 18 = O'Connor (Note 5); 19 = Olejnik (Note 7); 20 = Piché, Michlin, Rubin, & Johnson (1975); 21 = Rotenberg (1974); 22 = Rothenberg (1970); 23 = Rubin (1972); 24 = Rubin (1973); 25 = Rubin (1974); 26 = Rubin (1978); 27 = Rubin & Maioni (1975); 28 = Rubin & Schneider (1973); 29 = Selman (1971); 30 = Selman & Lieberman (1975); 31 = Sullivan & Hunt (1967); 32 = Turnure (1975); 33 = Urberg & Docherty (1976); 34 = Weinheimer (1972); 35 = Wolfe (1963); 36 = Zahn-Waxler, Radke-Yarrow, & Brady-Smith (1977). In most cases, no partial r was computed, so it does not appear in the entry. Some studies reported a range of correlations, and most used several separate age groups. IA is expressed as a percentage rather than as a correlation.

* $p < .05$, except for IA. Studies that reported only whether a correlation was significant or nonsignificant are denoted *sig* or *ns*.

Another study that found a modest but significant relation between the affective and cognitive/communicative domains (Moir, 1974) reported a correlation of .49 between Rothenberg's task and a game strategy inference measure. With IQ partialed out, the correlation was still significant ($r = .35$). One possible interpretation of these relationships is that some of these measures may be tapping some common social cognitive or personality dimension such as social insight or social sensitivity. However, without more validity information this hypothesis is little more than speculation. The alternative hypothesis that performance on perspective-taking measures in the affective and cognitive/communicative domains can be explained by an underlying construct of egocentrism is, as it was for the visual/spatial and affective domains, untenable.

Visual/Spatial Egocentrism and Cognitive/Communicative Egocentrism

A set of three studies that related visual/spatial and referential communication measures shows a fairly consistent pattern of significant correlations. Rubin (1973) obtained a correlation of .65 between Flavell's four displays and the Glucksberg-Krauss task on a sample of 80 children aged 5–12. With IQ partialed out, the correlation was considerably lower but still significant ($r = .35$). In another study of 112 children in grades two and six and college-aged and elderly adults, Rubin (1974) found a correlation of .49 between the same two tasks, although the correlation within ages was only significant for the sixth graders and undergraduates. Heilbrunn (1974) found a significant correlation for 8-year-olds but not for 12-year-olds between performance on a modified version of the three-mountains task and a referential communication task. These studies together suggest that there may be a common underlying construct that weakly ties the two sets of measures together.

Do other cognitive/communicative egocentrism measures show a similar relationship with visual/spatial egocentrism? With one possible exception, the answer seems to be no. Sullivan and Hunt (1967) found correla-

Table 2
Construct Validity and Reliability of Measures of Egocentrism

Measure	Feffer's Role-Taking Task	Privileged information	Infer game strategies
Feffer's Role-Taking Task	TR = .60 [N = 48, 6–10(13)] IC = .42 [N = 80, 6–13(7)] IC = .40 [N = 48, 6–10(13)] IC = .40 [N = 60, 7–12(32)] IC = .27 [N = 67, 12–13(11)] IR = .99 [N = 160, 4–10(16)] IR = .97 [N = 66, 7–12(32)] IR = .89 [N = 35, adults(6)] IR = .86 [N = 160, 5–8(34)] IR = .86 [N = 20, 9–10(20)] IR = .84 [N = 21, 10–21(35)] IA = 95% [N = 48, 6–10(13)]		
Privileged information	.23* (.06) [N = 96, 6–10(13)] .15 [N = 20, 9–10(20)]	TR = .84 [N = 125, 8–15(4)] TR = .68 [N = 48, 6–10(13)] IC = .91 [N = 125, 8–15(4)] IC = .65–.86 [N = 120, 5–9(19)] IC = .56 [N = 48, 6–10(13)] IR = .96 [N = 125, 8–15(4)] IR = .95 [N = 120, 5–9(19)] IR = .94 [N = 86, 6–12(3)] IR = .84 [N = 68, 9–12(15)] IR = .78 [N = 20, 9–10(20)] IA = 100% [N = 110, 7–8(10)] IA = 97% [N = 42, 3–5(32)] IA = 92% [N = 48, 6–10(13)]	
Infer game strategies	.34* (.22*) [N = 96, 6–10(13)]	.45* (.23*) [N = 96, 6–10(13)] (−.03) [N = 114, 3–11(26)]	TR = .41 [N = 48, 6–10(13)] IC = .77–.85 [N = 75, 3–5(18)] IC = .68 [N = 48, 6–10(13)] IA = 96% [N = 48, 6–10(13)] IA = 90% [N = 64, 10–adult(2)] IA = 100% [N = 12, 3–11(26)]

Note. TR = test–retest reliability; IC = internal consistency; IR = interrater reliability; IA = interrater agreement. For an explanation of the form of each entry, see Table 1 *Note.*
* $p < .05$, except for IA.

Table 3
Construct Validity and Reliability of Measures of Egocentrism

Measure	Private speech	Other conceptual role-taking tasks	Feffer's Role-Taking Task	Privileged information
Private speech	TR = .85–.95 [*N* = 10, 5–12 (24)] TR = .43 [*N* = 26, 4–7 (12)] IR = .90 [*N* = 26, 4–7 (12)] IR = .85 [*N* = 28, 4–7 (12)] IR = .82–.95 [*N* = 20, 5–12 (24)]			
Social speech	.68* [*N* = 28, 4–7 (12)]			
Other conceptual role-taking tasks		TR = .66 [*N* = 48, 6–10 (13)] TR = .62 [*N* = 68, 7 (30)] IC = .90–.93 [*N* = 108, 37–(36)] IC = .62 [*N* = 48, 6–10 (13)] IA = 99% [*N* = 48, 6–10 (13)] IA = 86–100% [*N* = 108, 3–7 (36)]	.49* (.35*) [*N* = 96, 6–10 (13)] .27 [*N* = 20, 9–10 (20)]	.38* (.09) [*N* = 96, 6–10 (13)] .12 [*N* = 20, 9–10 (20)]
IQ or mental age	.43* [*N* = 80, 5–12 (24)] .40* [*N* = 26, 4–5 (12)] −.04 [*N* = 26, 6–7 (12)]		*sig* [*N* = 68, 6–13 (7)] *sig* for some ages [*N* = 66, 7–12 (32)] *ns* [*N* = 36, adults (8)] .09–.24 [*N* = 90, 7–11 (31)]	.09 [*N* = 110, 7–8 (10)]
Conservation	.25* (−.04) [*N* = 80, 5–12 (24)]			
Popularity	.30 (.25) [*N* = 28, 4–7 (12)] −.02 [*N* = 80, 5–12 (24)]			

Note. TR = test–retest reliability; IC = internal consistency; IR = interrater reliability; IA = interrater agreement. For an explanation of the form of each entry, see Table 1 *Note.*
* $p < .05$, except for IA. Studies that reported only whether a correlation was significant or nonsignificant are denoted *sig* or *ns*.

tions of .25, .00, and .35 between the three-mountains task and Feffer's RTT at ages 7, 9, and 11, respectively. Heilbrunn similarly found no relationship at age 8 and a significant but modest correlation at age 12. Considering the poor internal consistency of the Feffer measure, it is not surprising that these correlations are unimpressive. The only study that correlated visual/spatial and privileged information tasks, Kurdek and Rodgon's (1975) study of 167 children from kindergarten to sixth grade, found nonsignificant correlations at all ages except grade five, which may be attributable to chance, considering the number of correlations computed. Moir (1974) found that one of two visual/spatial tasks was significantly correlated with Selman's measure of inferring game strategies, but Selman (1971b), using a sample of 60 preadolescents, did not obtain a significant relationship between his measure and two scores from a different visual/spatial perspective task. Rubin (1973) found a nonsignificant correlation of .28 (with IQ partialed out, −.06) between private speech and visual/spatial egocentrism. It seems that the only egocentrism measure other than referential communication that might be significantly related to visual/spatial perspective taking is comprehension of recursive thought; Rubin (1973) found a correlation of .73 (with IQ partialed out, a still significant .36) between these two measures.

Within-Domain Correlations for Cognitive/Communicative Egocentrism

Correlations among the various cognitive/communicative egocentrism tasks support the hypothesis that referential communication and recursive thought, along with visual/spatial perspective taking, form a cluster of measures that share some reliable variance beyond general intelligence and are more or less unrelated to other measures in this domain. Rubin (1973) found a correlation between measures of referential communication and recursive thought of .72, which remained significant when IQ was partialed out ($r = .31$). In Rubin's (1978) study, this finding was replicated (partial $r = .46$). In these two studies Rubin also found that neither measure was consistently related to private speech, the hide-the-penny game or Chandler's privileged information stories. In a sample of 20 fourth graders, Piché et al. (1975) also found nonsignificant correlations between Feffer's RTT and two referential communication tasks ($r = .23$ and $-.08$, respectively). However, one of these latter two measures, the Crystal Climbers Task, was significantly related to Chandler's stories ($r = .44$), although the other, the Baldwin and Garvey Picture Identification Task, was not ($r = .32$). Similarly, Chandler et al. (1974) found a modest but significant correlation of .31 between the privileged information stories and a referential communication task, whereas Leahy and Huard (1976) obtained a correlation of only .05. None of the above three studies controlled for age or IQ, which probably accounts for some of the small amount of shared variance between these two sets of measures. For example, Rubin (1978) obtained a significant *negative* correlation ($r = -.25$) between the Chandler and Glucksberg-Krauss tasks when chronological age was partialed out. Chandler's stories also do not consistently relate to Feffer's RTT (Kurdek, 1977; Piché et al., 1975), although low but significant correlations are reported by Kurdek between Chandler's stories and Flavell's nickel–dime game and between Feffer's RTT and the nickel–dime game. An important and revealing exception to this general pattern of low correlations among cognitive/communicative egocentrism measures is the Kohlberg et al. (1968) finding that private and social speech are highly correlated ($r = .68$). These measures probably reflect a general level of maturity in the use of language.

Interpretation and a Hypothesis

From these data one can argue that if there is a construct of egocentrism underlying performance on these tasks, it is not manifested on other than the visual/spatial, referential communication, and recursive thought measures. Even this conclusion is tentative, since it rests heavily on one study, and much of the shared variance among these tasks is attributable to general intelli-

gence. One might note that Rubin's (1973) factor analysis of these three measures, private speech, conservation, popularity, and chronological and mental age yielded two factors that are easily identifiable as general intelligence and popularity factors despite the fact that Rubin labeled the first factor an egocentrism factor. (This factor had high positive loadings from all variables except popularity.) Still, the modest but perhaps unexpected relationship among visual/spatial, referential communication, and recursive thought measures, if replicable, must be accounted for. Although egocentrism is one possible explanation, it is unparsimonious, since it fails to account for the predominantly nonsignificant correlations between these three tasks and other cognitive/communicative (as well as affective) egocentrism measures.

One can entertain the hypothesis that these measures are related beyond their simple relationship with general intelligence primarily because they share a property usually identified with a "purely" cognitive factor. This factor, labeled differently by different theorists (e.g., Bernyer, 1958; Horn, 1968; Vernon, 1965), may be interpreted at its most general level as involving the ability for spatial thinking and reasoning and general perceptual mastery. This hypothesis may seem more justifiable after examining the data on visual/spatial tasks and then analyzing the specific task requirements of the referential communication and recursive thought measures.

Further Consideration of Visual/Spatial Egocentrism Measures and Their Correlates

Several researchers have carefully analyzed the specific kinds of errors made by children on visual/spatial egocentrism measures, and one may refer to these for a more complete understanding of performance on these tasks (e.g., Coie et al., 1973; Eliot & Dayton, 1976; Fishbein et al., 1972; Huttenlocher & Presson, 1973; Shantz & Watson, 1971). The important point for this analysis is that in general, the proportion of egocentric errors is small at all ages, and the major developmental trend is not a tendency to make proportionately fewer egocentric errors but rather simply to make fewer errors of all kinds. Eliot and Dayton (1976), using a sample of 410 first graders, 421 fifth graders, and 260 adults and an adaptation of the three-mountains problem, found that egocentric (frontal) errors were made for the 90°, 180°, and 270° positions on only 16% of the trials with first graders and 8% of the trials with fifth graders. These proportions are not significantly different from each other. There was a highly significant decrease in the actual number of total errors made, however, and as expected, adults were proficient on this task. The authors concluded that young children are less perceptually accurate but not more perceptually egocentric.

Coie et al. (1973) found a similar tendency for egocentric errors to be relatively infrequent in a sample of 90 second, third, and fourth graders, although performance was far from perfect: Errors totaled 476, but only 80 of these were egocentric errors. They analyzed their data by ability levels (defined by performance on the egocentrism task) and did find a moderately significant trend towards making fewer egocentric errors at higher ability levels. Although this is more compatible with the observations reported by Piaget and Inhelder (1956) and Shantz and Watson (1971), these age or ability trends are easily overestimated. Coie et al. (1973) found that 20% of the errors in their lowest ability group were egocentric, which compares to a chance level of 8%. The authors concluded that development in this domain can best be characterized by small, undramatic transitions in the mastery of one's visual field.

A study even more striking in its demonstration of the small role played by egocentrism in the development of visual/spatial perspective-taking ability is a study by Fishbein et al. (1972). These researchers used a set of displays of increasing complexity briefly described in the last section (one or three toys and turning the display vs. pointing to one of either four or eight photographs). The complexity of the display and the mode of responding had significant effects on task performance, but the proportion of

egocentric errors did not show any tendency to decline with age. In fact, on the pointing task, preschoolers were less likely to make egocentric errors than were first graders across all variations of the display, even though the preschoolers made more total errors. Similarly, third graders were at least as likely as first graders to make egocentric errors. Together, these studies suggest that little of the variance in visual/spatial perspective-taking tasks can be attributed to egocentrism and that a straightforward cognitive reinterpretation might be tenable. Unfortunately, no study has related performance on visual/spatial tasks with scores on psychometric tests of spatial or perceptual thinking or with measures of spatial or perceptual processes such as mental rotation; consequently, this hypothesis requires further testing before confident conclusions can be drawn.

How do referential communication and recursive thought measures fit into this picture? A speculative hypothesis based on a simple analysis of the requirements of these tasks is that a major component of successful performance may be the ability to perceptually encode and discriminate the task stimuli. Referential communication measures such as that created by Glucksberg and Krauss (and used by Rubin in his 1973 study) use stimuli that are novel and difficult for children to perceptually process. The cognitive interpretation advanced here is supported by findings reported in Glucksberg et al. (1966). Up to age 4 (younger than Rubin's subjects) children were unable to successfully perform the task, even when pictures of familiar animals were substituted for the novel forms. However, between 52 and 63 months of age, all subjects could do the task with animals used as stimuli, but none could do it with the novel forms. Older subjects were not tested in this study, but another study by Hoy (1975) extends these observations of gradual improvement with age depending on the perceptual complexity and familiarity of the stimuli used as referents and nonreferents. Using 36 children aged 5, 7, and 9, Hoy found that children of all ages were better able to describe how to build a toy horse than they were a random shape and that performance improved with age for both stimuli. Similarly, Grushcow and Gauthier (1971) reported that 24 5-year-olds were more successful on a referential communication task using familiar rather than unfamiliar animals (67% vs. 46%) but were also more successful when unfamiliar animals were used than when familiar symbols were used (46% vs. 33%). When the task stimuli are chosen such that they are familiar and easily discriminable, children as young as 3 years old demonstrate an appreciation for the perspective of another: Maratsos (1973) found that children at this age were far more explicit verbally when communicating about the characteristics of simple referents (familiar toys) to an experimenter who apparently could not see than to an experimenter who could see.

In the recursive thought task, although subjects are taught to discriminate thinking cartoon clouds from talking clouds, most items are unusual and fairly complex perceptually. For example, if a boy is thinking of a girl who is thinking of the mother, three progressively smaller drawings of three different people appear in clouds of various shapes and sizes embedded in each other. Items of greater complexity may be even more difficult to decipher perceptually.

One important implication of this discussion is that although an awareness that others can have a different perspective and that one must infer that this perspective does seem to be required in most egocentrism measures, these abilities may be present in their simplest forms early developmentally (Borke, 1972). Although inferring perspective-taking ability is difficult in some egocentrism measures, presumably nonegocentric performance has been observed for children as young as 2 years old in the visual/spatial domain (Verkozen, 1975) and 3 years old in the affective (Borke, 1971) and cognitive/communicative domains (Maratsos, 1973; Menig-Peterson, 1975).

The implication of these findings for determining the sources of variance in performance on egocentrism tasks is that perspective-taking ability may account for little of the variance after age 4 or 5 (Mossler, Marvin, & Greenberg, 1976). More plausible sources of variance include (a) general intelligence

(Rubin, 1973); (b) verbal comprehension (Shantz & Watson, 1971); (c) specific cognitive factors such as spatial or perceptual abilities, depending on the complexity and/or familiarity of the task stimuli (Coie et al., 1973; Eliot & Dayton, 1976; Pufall, 1975); (d) characteristics specific to the type of response required (e.g., verbal vs. nonverbal; symbolic vs. concrete), since it may be the case that young children are unable to express their knowledge through the required response mode (Fishbein et al., 1972; Garber, 1975; Shantz, 1975); and (e) variables highly specific to the task, such as whether a real person or a doll is sitting in the position in which a visual/spatial perspective must be inferred, which Cox (1975) found made a significant difference (easier if a real person is used).

Summary. To summarize the data in this section and the implications of these data, visual/spatial, affective, and cognitive/communicative perspective-taking measures do not appear to tap a single unitary dimension of egocentrism. Measures of egocentrism are typically as highly correlated with other constructs (i.e., IQ, conservation, and popularity) as they are with measures of the same construct. Social speech and private speech measures may be tapping some common dimension that represents the developing child's mastery of language skills. Feffer's RTT and the measures of game strategy inference and awareness of privileged information seem to be independent of other measures in these domains, at least beyond any commonality due to general intelligence. For the most part, the same can be said for measures of affective egocentrism. This does not necessarily mean that these tests aren't measuring something that is consistent and important, since measurement reliabilities are generally adequate.

Visual/spatial, referential communication, and recursive thought measures seem to share some modest but significant amount of variance in addition to that attributable to general intelligence. Given that most errors made on visual/spatial perspective-taking tasks are not egocentric errors and that developmental improvements on these tasks are not characterized by dramatically smaller proportions of egocentric errors, the relationships among these three measures are more likely due to some common factor other than egocentrism, such as perceptual and/or spatial facility. For all of the measures that purport to measure egocentrism, task-specific and response-specific characteristics may account for a large proportion of the variance in performance, although this variance should not necessarily be construed as error, particularly if this variance can contribute to our knowledge about social cognitive development.

Conclusion

As a whole, the evidence reviewed in the last two sections and summarized in Tables 1–3 fails to support the construct validity of egocentrism. The most common finding is a lack of relationship among egocentrism measures, even for those whose reliability indicates that something consistent is being measured. The few commonalities found among specific tasks are more parsimoniously interpreted as the result of other explanatory constructs, such as those referring to the general level of cognitive, perceptual, or linguistic development of the child. Cronbach and Meehl (1955) comment that

> one who claims that his test reflects a construct cannot maintain his claim in the face of recurrent negative results because these results show that his construct is too loosely defined to yield verifiable inferences. (p. 291)

This statement is applicable to the research on egocentrism.

There are two main implications of negative evidence for construct validity. Either some or all of the tests are not good measures of the construct, or the theory that specifies the meaning of the construct is incorrect (Cronbach & Meehl, 1955). Since many of the egocentrism measures are reliable and possess a good deal of face validity, some doubt is cast on Piaget's theory, or at least that part of it that sets forth the meaning of egocentrism. Developing and testing alternative theoretical formulations that could better account for egocentric behavior is an investment of research effort that would most likely provide greater payoff in explanatory power.

Reference Notes

1. Baldwin, T. L., & Garvey, C. J. *Studies in convergent behavior: II. A measure of communication accuracy* (Report No. 91). Baltimore, Md.: Johns Hopkins University, Center for the Study of Social Organization of Schools, 1970.
2. Schnall, M., & Feffer, M. *Role-taking task scoring criteria*. Unpublished manuscript, 1960. (Available from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D.C. 20540. Order No. 9010. Remit $2.50 for microfilm or $6.25 for photocopies. Make checks payable to Chief of Photoduplication Service, Library of Congress.)
3. Ambron, S. R., & Irwin, D. M. *Role-taking and moral judgment in five- and seven-year-olds*. Paper presented at the meeting of the National Association for the Education of Young Children, Washington, D.C., November 1974.
4. Chandler, M. J., Helm, D., & Smith, M. *Developmental changes in the contribution of shared experience to social perspective taking skills*. Paper presented at a regional meeting of the Society for Research in Child Development, Chapel Hill, N.C., March 1974.
5. O'Connor, M. *Decentration revisited: A two-factor model for role-taking development in young children*. Paper presented at the meeting of the Society for Research in Child Development, Denver, Colo., April 1975.
6. Marsh, D., & Serafica, F. *Perspective taking and moral judgment: A developmental analysis*. Paper presented at the meeting of the Society for Research in Child Development, New Orleans, La., March 1977.
7. Olejnik, A. *Developmental changes and interrelationships among role-taking, moral judgments and children's sharing*. Paper presented at the meeting of the Society for Research in Child Development, Denver, Colo., April 1975.

References

Bernyer, G. Second order factors and the organization of cognitive functions. *British Journal of Statistical Psychology*, 1958, *11*, 19–29.

Borke, H. Interpersonal perception of young children: Egocentrism or empathy? *Developmental Psychology*, 1971, *5*, 263–269.

Borke, H. Chandler and Greenspan's "Ersatz egocentrism": A rejoinder. *Developmental Psychology*, 1972, *7*, 107–109.

Borke, H. The development of empathy in Chinese and American children between three and six years of age: A cross-culture study. *Developmental Psychology*, 1973, *9*, 102–108.

Burns, N., & Cavey, L. Age differences in empathic ability among children. *Canadian Journal of Psychology*, 1957, *11*, 227–230.

Byrne, D. E. The development of role-taking in adolescence (Doctoral dissertation, Harvard University, 1974). *Dissertation Abstracts International*, 1974, *34*, 5647B. (University Microfilms No. 74–11, 314)

Bunting, J. R. Egocentrism: The effects of social interaction through multi-age grouping (Doctoral dissertation, State University of New York at Buffalo, 1974). *Dissertation Abstracts International*, 1975, *35*, 6356A. (University Microfilms No. 74–29, 231)

Chandler, M. J. Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills. *Developmental Psychology*, 1973, *9*, 326–332.

Chandler, M. J., & Greenspan, S. Ersatz egocentrism: A reply to H. Borke. *Developmental Psychology*, 1972, *7*, 104–106.

Chandler, M. J., Greenspan, S., & Barenbaum, C. Assessment and training of role-taking and referential communication skills in institutionalized emotionally disturbed children. *Developmental Psychology*, 1974, *10*, 546–553.

Coie, J. D., Costanzo, P. R., & Farnill, D. Specific transitions in the development of spatial perspective-taking ability. *Developmental Psychology*, 1973, *9*, 167–177.

Cox, M. V. The other observer in a perspective task. *British Journal of Educational Psychology*, 1975, *45*, 83–85.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281–302.

Deutsch, F. Observational and sociometric measures of peer popularity and their relationship to egocentric communication in female preschoolers. *Developmental Psychology*, 1974, *10*, 745–747.

Eliot, J., & Dayton, C. M. Egocentric error and the construct of egocentrism. *Journal of Genetic Psychology*, 1976, *128*, 275–289.

Feffer, M. H. The cognitive implications of role taking behavior. *Journal of Personality*, 1959, *27*, 152–168.

Feffer, M. Developmental analysis of interpersonal behavior. *Psychological Review*, 1970, *77*, 197–214.

Feffer, M. H., & Gourevitch, V. Cognitive aspects of role-taking in children. *Journal of Personality*, 1960, *28*, 383–396.

Feffer, M., & Suchotliff, L. Decentering implications of social interactions. *Journal of Personality and Social Psychology*, 1966, *4*, 415–422.

Feshbach, N. D., & Roe, K. Empathy in six- and seven-year-olds. *Child Development*, 1968, *39*, 133–145.

Fishbein, H. D., Lewis, S., & Keiffer, K. Children's understanding of spatial relations: Coordination of perspectives. *Developmental Psychology*, 1972, *7*, 21–33.

Flavell, J. H., Botkin, P. T., Fry, C. L., Wright, J. W., & Jarvis, P. E. *The development of role-taking and communication skills in children*. New York: Wiley, 1968.

Garber, E. E. Development of the perception of spatial layout: The role of egocentrism (Doctoral dissertation, Cornell University, 1975). *Dissertation Abstracts International*, 1975, *36*, 2494B–2495B. (University Microfilms No. 75–24, 185)

Garvey, C., & Hogan, R. Social speech and social interaction: Egocentrism revisited. *Child Development*, 1973, *44*, 562–568.

Glucksberg, S., & Krauss, R. M. What do people say after they have learned how to talk? Studies of the development of referential communication. *Merrill-Palmer Quarterly*, 1967, *13*, 309–316.

Glucksberg, S., Krauss, R. M., & Higgins, R. The development of referential communication skills. In F. D. Horowitz (Ed.), *Review of child development research* (Vol. 4). Chicago: University of Chicago Press, 1975.

Glucksberg, S., Krauss, R. M., & Weisberg, R. Referential communication in nursery school children: Method and some preliminary findings. *Journal of Experimental Child Psychology*, 1966, *3*, 333–342.

Grushcow, R., & Gauthier, J. P. Effects of stimulus abstractness and familiarity on listener's performance in a communication task. *Child Development*, 1971, *42*, 956–958.

Heilbrunn, L. J. Egocentrism and the development of communication skills in children (Doctoral dissertation, State University of New York at Buffalo, 1973). *Dissertation Abstracts International*, 1974, *34*, 4020B. (University Microfilms No. 74–4404)

Horn, J. L. Organization of abilities and the development of intelligence. *Psychological Review*, 1968, *75*, 242–259.

Hoy, E. A. Measurement of egocentrism in children's communication. *Developmental Psychology*, 1975, *11*, 392.

Hudson, L. M. On the coherence of role-taking abilities: An alternative to correlational analysis. *Child Development*, 1978, *49*, 223–227.

Huttenlocher, J., & Presson, C. C. Mental rotation and the perspective problem. *Cognitive Psychology*, 1973, *4*, 277–299.

Keller, M. Development of role-taking ability: Social antecedents and consequences for school success. *Human Development*, 1976, *19*, 120–132.

Kohlberg, L., Yaeger, J., & Hjertholm, E. Private speech: Four studies and a review of theories. *Child Development*, 1968, *39*, 691–736.

Krauss, R. M., & Glucksberg, S. The development of communication: Competence as a function of age. *Child Development*, 1969, *40*, 255–266.

Kurdek, L. A. Structural components and intellectual correlates of cognitive perspective taking in first- through fourth-grade children. *Child Development*, 1977, *48*, 1503–1511.

Kurdek, L. A., & Rodgon, M. M. Perceptual, cognitive, and affective perspective taking in kindergarten through sixth-grade children. *Developmental Psychology*, 1975, *11*, 643–650.

Leahy, R. L., & Huard, C. Role taking and self-image disparity in children. *Developmental Psychology*, 1976, *12*, 504–508.

Liben, L. S. Perspective-taking skills in young children: Seeing the world through rose-colored glasses. *Developmental Psychology*, 1978, *14*, 87–92.

Looft, W. Egocentrism and social interaction across the life span. *Psychological Bulletin*, 1972, *78*, 73–92.

Maratsos, M. P. Nonegocentric communication abilities in preschool children. *Child Development*, 1973, *44*, 697–700.

Menig-Peterson, C. L. The modification of communicative behavior in preschool-aged children as a function of the listener's perspective. *Child Development*, 1975, *46*, 1015–1018.

Miller, P. H., Kessel, F. S., & Flavell, J. H. Thinking about people thinking about people thinking about . . .: A study of social cognitive development. *Child Development*, 1970, *41*, 613–623.

Moir, D. J. Egocentrism and the emergence of conventional morality in preadolescent girls. *Child Development*, 1974, *45*, 299–304.

Mossler, D. G., Marvin, R. S., & Greenberg, M. T. Conceptual perspective taking in 2- to 6-year-old children. *Developmental Psychology*, 1976, *12*, 85–86.

Piaget, J. *The language and thought of the child.* New York: Harcourt, Brace, 1926.

Piaget, J., & Inhelder, B. *The child's conception of space.* London: Routledge & Kegan Paul, 1956.

Piché, G. L., Michlin, M. L., Rubin, D. L., & Johnson, F. L. Relationships between fourth graders' performances on selected role-taking tasks and referential communication accuracy tasks. *Child Development*, 1975, *46*, 965–969.

Pufall, P. B. Egocentrism in spatial thinking: It depends on your point of view. *Developmental Psychology*, 1975, *11*, 297–303.

Rotenberg, M. Conceptual and methodological notes on affective and cognitive role taking (sympathy and empathy): An illustrative experiment with delinquent and nondelinquent boys. *Journal of Genetic Psychology*, 1974, *125*, 177–185.

Rothenberg, B. B. Children's social sensitivity and the relationship of interpersonal competence, intrapersonal comfort, and intellectual level. *Developmental Psychology*, 1970, *2*, 335–350.

Rubin, K. H. Relationship between egocentric communication and popularity among peers. *Developmental Psychology*, 1972, *7*, 364.

Rubin, K. H. Egocentrism in childhood: A unitary construct? *Child Development*, 1973, *44*, 102–110.

Rubin, K. H. The relationship between spatial and communicative egocentrism in children and young and old adults. *Journal of Genetic Psychology*, 1974, *125*, 295–301.

Rubin, K. H. Role taking in childhood: Some methodological considerations. *Child Development*, 1978, *49*, 428–433.

Rubin, K. H., & Maioni, T. L. Play preference and its relationship to egocentrism, popularity, and classification skills in preschoolers. *Merrill-Palmer Quarterly*, 1975, *21*, 171–179.

Rubin, K. H., & Schneider, F. W. The relationship between moral judgment, egocentrism, and altruistic behavior. *Child Development,* 1973, *44,* 661–665.

Selman, R. L. The relation of role taking to the development of moral judgment in children. *Child Development,* 1971, *42,* 79–91. (a)

Selman, R. L. Taking another's perspective: Role-taking development in early childhood. *Child Development,* 1971, *42,* 1721–1734. (b)

Selman, R. L., & Lieberman, M. Moral education in the primary grades: An evaluation of a developmental curriculum. *Journal of Educational Psychology,* 1975, *67,* 712–716.

Shantz, C. The development of social cognition. In E. M. Hetherington (Ed.), *Review of child development research* (Vol. 5). Chicago: University of Chicago Press, 1975.

Shantz, C. U., & Watson, J. S. Spatial abilities and spatial egocentrism in the young child. *Child Development,* 1971, *42,* 171–181.

Shatz, M., & Gelman, R. The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the Society for Research in Child Development,* 1973, *38*(5, Serial No. 152).

Sullivan, E. V., & Hunt, D. E. Interpersonal and objective decentering as a function of age and social class. *Journal of Genetic Psychology,* 1967, *110,* 199–210.

Turnure, C. Cognitive development and role-taking ability in boys and girls from 7 to 12. *Developmental Psychology,* 1975, *11,* 202–209.

Urberg, K. A., & Docherty, E. M. Development of role-taking skills in young children. *Developmental Psychology,* 1976, *12,* 198–203.

Verkozen, J. Egocentrism: Stage or state? *Psychoanalytic Review,* 1975, *62,* 305–308.

Vernon, P. E. Ability factors and environmental influences. *American Psychologist,* 1965, *20,* 723–733.

Weinheimer, S. Egocentrism and social influence in children. *Child Development,* 1972, *43,* 567–578.

Wolfe, R. The role of conceptual systems in cognitive functioning at varying levels of age and intelligence. *Journal of Personality,* 1963, *31,* 108–123.

Zahn-Waxler, C., Radke-Yarrow, M., & Brady-Smith, J. Perspective-taking and prosocial behavior. *Developmental Psychology,* 1977, *13,* 87–88.

Received July 24, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1189–1200

Large Contingency Tables With Large Cell Frequencies: A Model Search Algorithm and Alternative Measures of Fit

Douglas A. Zahn
Department of Statistics, Florida State University and Harvard University

Sara Beck Fein
Research for Social Change, Washington, D. C.

A new search algorithm, the Generalized Guided Method, for locating a model for contingency tables and a measure of fit related to the R^2 of multiple regression are proposed for the analysis of contingency tables with many cells and large cell frequencies. This algorithm is designed to analyze contingency tables that contain dichotomous and/or nondichotomous (polytomous) variables. A $4 \times 2 \times 15 \times 7 \times 2$ contingency table with 559,158 observations is used for illustration.

The user of conventional multidimensional contingency table techniques, such as those described by Goodman (1972a, 1972b), encounters two problems in the analysis of contingency tables with many cells and with large cell frequencies: (a) which measures of fit to use in assessing and comparing models and (b) how to search for appropriate models among the multitude of possible hierarchical models. This article, building primarily on the work of Goodman, addresses these problems. A model selection criterion and a measure of fit that is less dependent on sample size than the traditional chi-square measures are proposed, and a new search algorithm for locating models that satisfactorily fit large contingency tables is described.

The data used for illustration are from Miller, Simons, and Fein (1974) and involve characteristics of 559,158 persons admitted to British mental hospitals. Miller et al. were interested in whether the legal code under which a person is admitted to the hospital varies by age (four age categories), sex (two sexes), region (15 regions), and year of admission (7 years). The two categories of the variable Legal Code are formal (similar to involuntary commitment) and informal (similar to voluntary commitment).

Requests for reprints should be sent to Douglas A. Zahn, Department of Statistics, Florida State University, Tallahassee, Florida 32306.

Notation and Terminology

The symbols A, S, R, Y, and L denote the variables in the five-way, $4 \times 2 \times 15 \times 7 \times 2$ contingency table under consideration: age, sex, region, year, and legal code, respectively. Let $f_{ijk\ell m}$ denote the observed frequency in cell (i, j, k, ℓ, m). Let $F_{ijk\ell m}$ denote the expected value of this cell under the model being fitted, and let $\hat{F}_{ijk\ell m}$ denote the estimate of $F_{ijk\ell m}$. F and $\hat{F}$ may be superscripted to indicate a specific model.

Attention is restricted here to the situation in which one of the dichotomous variables in the contingency table is viewed as a dependent variable. These techniques may also be used in situations with polytomous dependent variables, although the results are more difficult to interpret, as explained in Goodman (1971). Considering L as the dependent variable, let $\omega_{ijk\ell} = f_{ijk\ell 1}/f_{ijk\ell 2}$ denote the observed odds in favor of formal admission for patients in cell (i, j, k, ℓ). Let $\Omega_{ijk\ell}$ denote the expected odds that the admission is formal under the model being considered, where $\Omega_{ijk\ell} = F_{ijk\ell 1}/F_{ijk\ell 2}$. Let $\Phi_{ijk\ell} = ln\, \Omega_{ijk\ell}$, where ln denotes the natural logarithm. Thus Φ denotes the logarithm of the expected odds that an

admission is formal and has been termed by Bishop, Fienberg, and Holland (1975) and others the logit pertaining to variable L. The saturated model for Φ_{ijkl} can be written

$$\Phi_{ijkl} = \beta + \beta_i^{A} + \beta_j^{S} + \beta_k^{R} + \beta_l^{Y} + \beta_{ij}^{AS} + \beta_{ik}^{AR} + \beta_{il}^{AY} + \beta_{jk}^{SR} + \beta_{jl}^{SY} + \beta_{kl}^{RY} + \beta_{ijk}^{ASR} + \beta_{ijl}^{ASY} + \beta_{ikl}^{ARY} + \beta_{jkl}^{SRY} + \beta_{ijkl}^{ASRY}, \quad (1)$$

where the βs satisfy the constraints

$$\sum_{i=1}^{4} \beta_i^{A} = 0, \ldots, \sum_{i=1}^{4} \beta_{ijkl}^{ASRY} = \sum_{j=1}^{2} \beta_{ijkl}^{ASRY} = \sum_{k=1}^{15} \beta_{ijkl}^{ASRY} = \sum_{l=1}^{7} \beta_{ijkl}^{ASRY} = 0. \quad (2)$$

The order of an effect in the logit model is the number of letters in its superscript. An effect may be denoted by the letters in its superscript; for example, AS may be used to denote β_{ij}^{AS}. One effect is said to be a lower order relative of a second effect if the letters in the superscript of the first are a subset of those in the superscript of the second. Thus the third-order effects ASR and ASY are lower order relatives of the fourth-order effect ASRY. A hierarchical model is one in which, for each interaction appearing in the model, all lower order relative effects are also in the model.

Attention is restricted here to hierarchical models. Goodman's (1972b) "minimal set of marginal tables fitted under the model" (p. 39) notation will be used to denote the models that are considered. However, this notation presents a problem, since it refers to cell frequency models, that is, to models for $ln\, F_{ijklm}$ rather than to logit models, such as the one given in Equation 1. The difficulty can be circumvented by determining for each logit model its equivalent cell frequency model. A cell frequency model and a logit model are said to be equivalent if estimates of the parameters of the logit model that use the estimated cell frequency model parameters are equal to those produced by direct estimation of the logit model parameters. Since in logit analyses the contingency table formed by the independent variables is assumed to be fixed, the equivalent cell frequency models must preserve this table (see Goodman, 1971). Where L is the dependent variable, the cell frequency models that are equivalent to the logit models being fitted must preserve the (ASRY) margin. Using Goodman's techniques and notation, it can be shown that the cell frequency model that is equivalent to the logit model

$$\Phi_{ijkl} = \beta + \beta_i^{A} + \beta_j^{S} + \beta_k^{R} + \beta_{ik}^{AR} \quad (3)$$

preserves the (ASRY), (SL), and (ARL) margins. This model can be identified by the abbreviated list of β parameters included in it, that is, the parameters that have no higher order relatives in this logit model, namely S and AR.

Another aspect of the equivalence of cell frequency and logit models is that they yield the same estimated cell frequencies, $\hat{F}$. Using the estimated logits, $\hat{\Phi}_{ijkl}$ and the entries in the (ASRY) margin, $f_{ijkl+} = f_{ijkl1} + f_{ijkl2}$, one can compute estimated cell frequencies using the relation

$$\hat{F}_{ijkl1} = (f_{ijkl+}) \exp(\hat{\Phi}_{ijkl}) / [1 + \exp(\hat{\Phi}_{ijkl})], \quad (4)$$

where $\exp(x) = e^x$. Also, the $\hat{F}$s from the cell frequency model can be used to compute estimated logits that are equivalent to those obtained using the logit models, by the formula $\hat{\Phi}_{ijkl} = ln\,(\hat{F}_{ijkl1}/\hat{F}_{ijkl2})$. Since the $\hat{F}$s produced by logit models are equivalent to those produced by cell frequency models, the fit of a logit model may be assessed by the likelihood-ratio chi-square goodness-of-fit statistic for the equivalent cell frequency model.

Model A is said to be nested in Model B if Model A is a special case of Model B that can be obtained by setting some parameter(s) in Model B equal to zero. For example, the model (ARL), (ASRY) is nested in the model (ASL), (ARL), (ASRY).

Measures of Fit for Contingency Tables With Large Cell Frequencies

The hypothesis H that a specific model fits the data in a contingency table is examined by first determining the maximum likelihood estimates, assuming H is true, of the expected cell frequencies in the table, by using the iterative proportional fitting procedure, which is also called the Deming-Stephan (1940) algorithm. This can be done with any of several

contingency table computer programs that are available (e.g., Dixon, 1975; Goodman, 1973; Haberman, 1972; Zahn, Note 1). The use of this algorithm is illustrated in several articles, including those by Davis (1974), Fienberg (1970), and Goodman (1972a). Conventionally, the differences between the observed cell frequencies f_{ijklm} and the estimated expected cell frequencies $\hat{F}_{ijklm}$ under hypothesis H are examined to determine if the discrepancies are large enough to cast doubt on the hypothesis that H fits the data, by using the likelihood ratio chi-square statistic

$$\chi^2(\mathrm{H}) = 2\sum_i \sum_j \sum_k \sum_l \sum_m [f_{ijklm} \times ln\,(f_{ijklm}/\hat{F}_{ijklm})]. \quad (5)$$

This statistic has an asymptotic chi-square distribution with degrees of freedom denoted $df(\mathrm{H})$. (For an extended discussion of the calculation of $df(\mathrm{H})$, see Bishop et al., 1975, section 3.8; and Davis, 1974, pp. 205–208, 213). The computer programs listed above provide $df(\mathrm{H})$.

Problems arise if $\chi^2(\mathrm{H})$ is used as a measure of fit in a contingency table with large cell frequencies. Bishop et al. (1975, section 9.6) show that the magnitude of this statistic is proportional to the sample size if the hypothesis H is not exactly true. With large cell frequencies, only the saturated model may yield an insignificant chi-square statistic. Terms of marginal utility may be incorporated into the final model simply because their chi-square statistics are inflated to significance by the large sample size. This sacrifices the parsimony often desired in describing a data set. In addition, analysis of models that contain terms that have small effects may yield less accurate cell frequency estimates than does analysis of simpler models (Bishop et al., 1975, section 9.2; see Hocking, 1976, for analogous multiple regression results). These arguments also speak against the suggestion that with a very large sample size, the only appropriate analysis, rather than hypothesis testing, is the estimation of parameters in the saturated model.

In developing an alternative to chi-square for assessing the goodness of fit of a model, the assessment of fit of multiple regression models, which has been more extensively studied, should be considered. Namboodiri, Carter, and Blalock (1975, p. 458), in a discussion of models based on regression methods, suggest that the researcher not rely solely on tests of significance in assessing fit (a) because large sample sizes would lead to rejecting models with adequate fit and (b) because in the social sciences, even with a perfect model, measurement errors in the variables may produce distortions great enough to cause the model to be rejected. Among the suggestions by Draper and Smith (1966, pp. 165, 238) for assessing fit are (a) the subjective examination of the increase in R^2 as additional variables are added, looking for "breaks" in the rate of increase, after which additional variables add little explanatory power; (b) the setting of an arbitrary proportion of the total variance that must be explained for the model to be considered adequate. In their example, an R^2 of .80 is selected. There is, then, precedence for moving away from tests of significance as criteria of fit in the development of models.

Both of Draper and Smith's (1966) suggestions for alternative measures of fit are based on the assessment of the proportion of variation explained by the model. Goodman (1971, p. 54; 1972a, p. 1057) proposed a statistic analogous to the R^2 of multiple regression for contingency tables. He defined the statistic

$$R^2 = [\chi^2(\mathrm{H}_0) - \chi^2(\mathrm{H})]/\chi^2(\mathrm{H}_0), \quad (6)$$

where H_0 denotes the hypothesis $\Phi_{ijkl} = \beta$, that is, that the logits are constant, and therefore the effects of all variables in the logit model are zero.

The just defined R^2 statistic simultaneously reflects two pieces of information about Model H: (a) R^2 itself measures the proportion of the total variation in the table, as indicated by the lack of fit for Model H_0, $\chi^2(\mathrm{H}_0)$, which is explained by Model H; and (b) $1 - R^2$ measures the proportion of variation still unexplained in the table. This partition relates to the two chi-square statistics that are of interest in assessing any contingency table model. The first is the difference $\chi^2(\mathrm{H}_0) - \chi^2(\mathrm{H})$, which tests the significance of the parameters in Model H, whereas the second, $\chi^2(\mathrm{H})$, tests the

lack of fit of Model H or the significance of the parameters in the saturated model that are not in Model H. The first of these is analogous to the multiple regression F test of the hypothesis that all the parameters in the regression model are equal to zero, whereas the second statistic does not have an analogue in multiple regression, since in the latter context we do not have a test statistic available for the determination of whether the variance unexplained is statistically significant.

Our use of R^2 goes beyond Goodman's (1972a) recommendation that this statistic be used as a measure of how well Model H fits the data. We propose that R^2 may be a better criterion for choosing among models for contingency tables with large cell counts than is chi-square.

Comments on R^2

Perhaps the most difficult question relating to the use of R^2 as an indicator of the fit of a given model is when a model can be said to fit adequately. If the researcher were depending on statistical significance as a criterion, a cutoff point could be chosen, though the choice of a specific significance level is difficult. Even in multiple regression models, which have a longer history of usage, traditional cutoff points for the adequacy of R^2 have not been developed. Also, recent work in multiple regression indicates that judging when a multiple regression equation fits adequately may be a more difficult question than previously thought (Mosteller & Tukey, 1977, chap. 12–16). An additional problem in contingency tables is that researchers have considerably less experience in determining R^2 values of the kind proposed here than they do for multiple regression studies.

In general, the choice encountered in adding more parameters is between increased explanation of variation, that is, increased R^2, and increased complexity of the model. Results in the regression literature, as summarized by Hocking (1976), imply that the more interactions with small effects there are in the model, the less precisely the logits are estimated. As parameters are added to the model, generally the rate of increase in R^2 per parameter added decreases rapidly. Thus there comes a point in most contingency table analyses at which additional increments in the R^2 statistic will cost dearly.

Perhaps initial insight into the question of adequate magnitudes of R^2 can be developed by an examination of Table 1, which reflects the magnitudes of R^2 statistics evident in contingency tables fitted in the literature using conventional criteria. Note that the R^2 values for models with statistically insignificant chi-square statistics range from .796 to .9996.

Table 1
Sample Size, Degrees of Freedom, χ^2*, and* R^2 *Values for Models That Were Assessed in Recent Literature as Fitting Contingency Tables*

Source	n	df	χ^2	R^2
Goodman (1971)	1,008	8	5.66	.827
Goodman (1972b)	8,036	2	1.32	.9996
Goodman (1973)				
Model 8	2,982	2	.31	.995
Model 16	2,982	10	6.11	.911
Model 20	2,982	14	13.95	.796

Additional Alternatives to Chi-Square

A number of alternatives to chi-square, other than R^2, have been proposed for assessing fit of contingency table models. One of these is χ^2/N, where N denotes the total number of counts in the contingency table. This statistic enables the researcher to compare the lack of fit of various models, either from the same or different tables. The statistic R^2 can also be used to do this, and in addition, it measures how much improvement the current model offers over H_0.

A second alternative measure of fit is the correlation between the actual type of admission for a subject and the predicted probability of formal admission. The models for $\Phi_{ijk\ell}$ can be used to predict probabilities of formal admission. The predicted probability of formal admission for subjects in cell (i, j, k, ℓ) under the model being fitted is

$$\exp(\hat{\Phi}_{ijk\ell})/[1 + \exp(\hat{\Phi}_{ijk\ell})] = \hat{F}_{ijk\ell 1}/(\hat{F}_{ijk\ell 1} + \hat{F}_{ijk\ell 2}). \quad (7)$$

However, the correlation between a dichotomous variable and a predicted probability may

well be small under reasonable models, as demonstrated by Morrison (1972), although Goldberger (1973) has indicated that such a correlation can reach the bound of +1.0.

Another potential measure of fit reflects the fact that $R^2 = 1$ does not indicate that the model predicts perfectly whether an individual has a formal or informal admission; what is perfectly predicted is the proportion of formal admissions for individuals in cell (i, j, k, ℓ) for all such cells. Hence, another possible measure of fit is the proportion of subjects correctly classified as formal or informal admissions.

The measure of fit that is appropriate depends on the objectives of the research study. However, the consequences of using the different measures of fit and which measure is best in what circumstances appear to be open questions.

Generalized Guided Method for Locating Models for Large Contingency Tables

The saturated logit model for the $4 \times 2 \times 15 \times 7 \times 2$ contingency table that is used as an example in this article illustrates one of the problems encountered with large contingency tables: the number of linearly independent parameters in the model is 840, far too many to interpret easily. Since the saturated logit model will always fit the data perfectly, whether a satisfactory fit can be obtained using a simpler model is a matter of importance. With a large contingency table, it is important to have a computationally feasible plan for searching among the multitude of possible models. With only four independent variables, there are 166 possible models. Several algorithms have been proposed for finding an adequate model (Bishop et al., 1975, Fienberg, 1970; Goodman, 1973; Shaffer, 1973).

The first step of Goodman's guided method is to compute the standardized effect estimates, which are used to construct a series of nested models. A major problem with this method is that the magnitude of an effect cannot be measured by a single effect estimate unless all variables in the effect are dichotomous. For example, there are 420 parameters in the Legal Code $\times$ Age $\times$ Region $\times$ Year effect. This makes it virtually impossible to use Goodman's procedure to determine the order of entry of the effects. One reason for developing the method described below is to provide a procedure for summarizing the relative importance of the various effects in an interpretable way.

Shaffer (1973) suggests an approach for the determination of which parameters are important, but her approach is not feasible when large tables are analyzed. Fienberg's (1970) procedure is based on a posited series of nested hierarchical models but requires that the contingency table be well enough understood to construct the series of models before examining the data.

Bishop et al. (1975, section 4.5) discuss several search algorithms and emphasize that the strategy used should be sensitive to the specific research problem. The algorithm proposed here uses some of their recommendations, deletes others that in our experience have not been practical in consulting problems that involve contingency tables, and includes additional modifications we have found useful.

Higgins and Koch (1977) describe another approach to the analysis of large contingency tables that uses Pearson chi-square statistics, Mantel-Haenszel statistics, and the model-fitting procedure developed by Grizzle, Starmer, and Koch (1969). Benedetti and Brown (1978) also discuss several search algorithms. They stress the importance of examining each effect in the model, as is done in the algorithm described below, which is similar to the algorithm they propose.

Our search for a solution to the problems noted in previous methods has led us to develop the Generalized Guided Method (GGM), patterned after a method recommended for locating regression models when the number of independent variables is large. In these situations, some authors (e.g., Daniel & Wood, 1971) have recommended fitting a regression model by using all available variables and computing the partial F statistics for all variables in the model. For a given variable, the numerator of this statistic is the sum of squares explained by that variable after all other variables have been entered into the model. This is almost always a conservative measure of the explanatory power of this variable, since generally, if this variable were not the last to enter, it would explain more

Table 2
Statistics for Pairs of Models to Select Entry Order of Effects in Models for the Five-Way Table

Abbreviated list of β parameters included in logit model	Fitted marginals	β parameter under consideration	df	χ^2	$C(\beta)$	Entry order
SRY SRY, A	(ASRY), (SRYL) (ASRY), (SRYL), (AL)	A	630 627	2,515.38 1,927.72	195.89	1
ARY ARY, S	(ASRY), (ARYL) (ASRY), (ARYL), (SL)	S	420 419	1,231.36 1,148.22	83.14	5
ASY ASY, R	(ASRY), (ASYL) (ASRY), (ASYL), (RL)	R	784 770	3,752.30 2,185.31	111.93	4
ASR ASR, Y	(ASRY), (ASRL) (ASRY), (ASRL), (YL)	Y	720 714	2,495.43 1,768.01	121.24	3
ARY, SRY ARY, SRY, AS	(ASRY), (ARYL), (SRYL) (ASRY), (ARYL), (SRYL), (ASL)	AS	315 312	1,014.37 420.88	194.50	2
ASY, SRY ASY, SRY, AR	(ASRY), (ASYL), (SRYL) (ASRY), (ASYL), (SRYL), (ARL)	AR	588 546	1,196.15 807.08	9.26	7
ASR, SRY ASR, SRY, AY	(ASRY), (ASRL), (SRYL) (ASRY), (ASRL), (SRYL), (AYL)	AY	540 522	840.74 716.62	6.90	8
ASY, ARY ASY, ARY, SR	(ASRY), (ASYL), (ARYL) (ASRY), (ASYL), (ARYL), (SRL)	SR	392 378	537.57 474.97	3.76	9
ASR, ARY ASR, ARY, SY	(ASRY), (ASRL), (ARYL) (ASRY), (ASRL), (ARYL), (SYL)	SY	360 354	389.87 386.71	.53	14
ASR, ASY ASR, ASY, RY	(ASRY), (ASRL), (ASYL) (ASRY), (ASRL), (ASYL), (RYL)	RY	672 588	1,625.55 766.29	10.23	6

Table 2 (*Continued*)

Abbreviated list of β parameters included in logit model	Fitted marginals	β parameter under consideration	df	χ^2	$C(\beta)$	Entry order
ASY, ARY, SRY ASY, ARY, SRY, ASR	(ASRY), (ASYL), (ARYL), (SRYL) (ASRY), (ASYL), (ARYL), (SRYL), (ASRL)	ASR	294 252	402.91 296.22	2.54	10
ASR, ARY, SRY ASR, ARY, SRY, ASY	(ASRY), (ASRL), (ARYL), (SRYL) (ASRY), (ASRL), (ARYL), (SRYL), (ASYL)	ASY	270 252	314.45 296.22	1.01	12
ASR, ASY, SRY ASR, ASY, SRY, ARY	(ASRY), (ASRL), (ASYL), (SRYL) (ASRY), (ASRL), (ASYL), (SRYL), (ARYL)	ARY	504 252	698.25 296.22	1.60	11
ASR, ASY, ARY ASR, ASY, ARY, SRY	(ASRY), (ASRL), (ASYL), (ARYL) (ASRY), (ASRL), (ASYL), (ARYL), (SRYL)	SRY	336 252	368.38 296.22	.86	13

variation. In searching for the final model in these regression situations, attention is focused on those variables with the largest partial F statistics.

In the first step of the GGM, the researcher computes a conservative measure of the explanatory power of each effect being considered for inclusion in the model. This measure is based on the reduction in the chi-square statistic due to adding that effect to the model after as many effects as possible (preserving the hierarchical nature of the model) have been included. Computation of this measure requires fitting a pair of models, denoted H* and H**, for each effect. The first model, H*, excludes only the effect of interest, denoted β^*, and its higher order relatives; it includes all other effects. Model H** includes the effects in H* and the effect of interest. Thus $\chi^2(\mathrm{H}^*) - \chi^2(\mathrm{H}^{**})$ can be used to test the hypothesis that the effect of interest is equal to zero, when as many effects as possible are entered into the model before it (cf. Goodman, 1972a, pp. 1051–1052).

For example, for the effect β_{ij}^{AS}, Models H* and H** are, respectively,

$$\Phi_{ijk\ell}=\beta+\beta_i^{\mathrm{A}}+\beta_j^{\mathrm{S}}+\beta_k^{\mathrm{R}}+\beta_\ell^{\mathrm{Y}}+\beta_{ik}^{\mathrm{AR}}+\beta_{i\ell}^{\mathrm{AY}} +\beta_{jk}^{\mathrm{SR}}+\beta_{j\ell}^{\mathrm{SY}}+\beta_{k\ell}^{\mathrm{RY}}+\beta_{ik\ell}^{\mathrm{ARY}}+\beta_{jk\ell}^{\mathrm{SRY}},$$

and

$$\Phi_{ijk\ell}=\beta+\beta_i^{\mathrm{A}}+\beta_j^{\mathrm{S}}+\beta_k^{\mathrm{R}}+\beta_\ell^{\mathrm{Y}}+\beta_{ij}^{\mathrm{AS}}+\beta_{ik}^{\mathrm{AR}} +\beta_{i\ell}^{\mathrm{AY}}+\beta_{jk}^{\mathrm{SR}}+\beta_{j\ell}^{\mathrm{SY}}+\beta_{k\ell}^{\mathrm{RY}}+\beta_{ik\ell}^{\mathrm{ARY}}+\beta_{jk\ell}^{\mathrm{SRY}}. \quad (8)$$

Thus this measure of the explanatory power of β_{ij}^{AS} is $\chi^2(\mathrm{H}^*)-\chi^2(\mathrm{H}^{**})=1{,}014.37-420.88=583.49$.

The difference $\chi^2(\mathrm{H}^*) - \chi^2(\mathrm{H}^{**})$ gives a misleading ranking of effects, since the numbers of linearly independent parameters associated with the effects are unequal. To adjust for this, the statistic $C(\beta^*)$ is used in the first step of the GGM to order the effects according to their explanatory power per linearly independent parameter, where

$$C(\beta^*) = [\chi^2(\mathrm{H}^*) - \chi^2(\mathrm{H}^{**})]/ [df(\mathrm{H}^*) - df(\mathrm{H}^{**})]. \quad (9)$$

For the effect β^{AS}, we obtain $C(\beta_{ij}^{\mathrm{AS}}) = (1{,}014.37 - 420.88)/(315 - 312) = 194.50$. Thus if effect β_{ij}^{AS} is added to the Model H*, the lack of fit is reduced by 194.50 per linearly independent parameter added. The addition of the effect β_{ij}^{AS} to the model actually

Table 3
χ^2 and R^2 Statistics for a Nested Series of Models for the Five-Way Table

Model	Abbreviated list of β parameters included in logit model	Fitted marginals	β parameter added to previous logit model	*df*	χ^2	R^2
H_0	Grand *M*	(ASRY), (L)		839	5,932.93	.00
H_1	A	(ASRY), (LA)	A	836	5,346.96	.09
H_2	AS	(ASRY), (LAS)	AS	832	4,664.64	.21
H_3	AS, Y	(ASRY), (LAS), (LY)	Y	826	3,894.85	.34
H_4	AS, Y, R	(ASRY), (LAS), (LY), (LR)	R	812	2,326.62	.60
H_5	AS, RY	(ASRY), (LAS), (LRY)	RY	728	1,479.77	.75
H_6	AS, RY, AR	(ASRY), (LAS), (LRY), (LAR)	AR	686	1,093.03	.81
H_7	AS, RY, AR, AY	(ASRY), (LAS), (LRY), (LAR), (LAY)	AY	668	968.83	.83
H_8	AS, RY, AR, AY, SR	(ASRY), (LAS), (LRY), (LAR), (LAY), (LSR)	SR	654	906.62	.84
H_9	RY, AY, ASR	(ASRY), (LRY), (LAY), (LASR)	ASR	612	798.36	.86
H_{10}	ASR, ARY	(ASRY), (LASR), (LARY)	ARY	360	389.87	.93
H_{11}	ASR, ARY, ASY	(ASRY), (LASR), (LARY), (LASY)	ASY	336	368.38	.93
H_{12}	ASR, ARY, ASY, SRY	(ASRY), (LASR), (LARY), (LASY), (LSRY)	SRY	252	296.22	.95

Note. This series is derived from results in Table 2.

involves including eight more parameters, three of which are linearly independent.

The results obtained in the first step of the GGM for our example are presented in Table 2. This step involves fitting 25 separate models.

In the second step of the GGM, a series of hierarchical models is constructed, starting with H_0 and adding effects one at a time in the order of the $C(\beta)$ statistics. The order of entry of the effects into the series of models is indicated in the last column of Table 2. For our example, this step produced the series of 12 unsaturated models listed in Table 3.

The third step in the algorithm is to fit the 12 models in the series. The results are presented in Table 3. At this point in the process, the researcher may use several strategies to select a reasonable model. We suggest that the researcher consider the following criteria (cf. Draper & Smith, 1966, pp. 165, 458): (a) The R^2 statistics in the series of nested models might be examined for "breaks," or points at which further increases in R^2 require a large number of additional parameters, and/or points at which the R^2 values level off. Using this criterion, either Model H_6 or Model H_{10} might be chosen from Table 3. (b) The researcher might decide what proportion of the total variation in the logits in the table must be explained for a model to be deemed adequate. Let R_T^2 denote this proportion. Then the simplest model H_u in the nested series in which $R^2 \geq R_T^2$ is identified. Model H_u and all other models in which $R^2 \geq R_T^2$ will be said to be "R_T^2 adequate."

A goal of $R_T^2 = .80$ has been chosen here. An examination of Table 3 indicates that $H_u = H_6$; that is, H_6 is the simplest model in the nested series that is .80 adequate. We note that there are other models for this table that are .80 adequate, including $H_7, H_8, \ldots, H_{12}$. Unfortunately, identifying all .80-adequate models requires fitting many models to this contingency table. We recommend this step when it is computationally feasible. However, when the number of dimensions is large, this procedure is expensive; it is for these situations that the GGM is useful. We fitted all possible logit models to establish a standard for evaluating the GGM. This step located 40 models with $R^2 \geq .80$.

The fourth and final step of the GGM is to identify a minimal R^2_T-adequate model. Model H will be termed minimal R^2_T adequate (a) if $R^2(\mathrm{H}) \geq R^2_T$ and (b) if for any Model H′ that can be constructed by deleting one or more effects from H, preserving the hierarchical nature of the model, $R^2(\mathrm{H}') < R^2_T$. This criterion reduces the 40 models that are .80 adequate to the following three minimal .80-adequate models: (a) (ASRY), (ARL), (RYL), (AYL), (SYL); (b) (ASRY), (ARYL), (SL); (c) (ASRY), (ASL), (RYL), (ARL) (Model H_6).

For illustration, the calculations are presented in Table 4 for Model H_6 only, the model found through earlier steps of the GGM. These calculations indicate that the removal of any effect from Model H_6 yields $R^2 < .80$; hence, Model H_6 is minimal .80 adequate.

In summary, we note that the GGM located one of the three minimal .80-adequate models for this table. The selected model is

$$\text{Model } H_6: \Phi_{ijkl}=\beta+\beta_i^{A}+\beta_j^{S}+\beta_k^{R}+\beta_l^{Y}+\beta_{ij}^{AS}+\beta_{ik}^{AR}+\beta_{kl}^{RY}. \quad (10)$$

At this point, the researcher knows that no three-way interactions are necessary to explain at least 80% of the variation in the logits; with reasonable accuracy the odds ratio of the number of informal to the number of formal admissions for any cell in the Age × Sex × Region × Year four-way table of logits can be predicted using only the parameters indicated in Model H_6. Thus instead of having to contend with one four-dimensional table with 840 parameters, the researcher knows that three two-way tables, with 8, 60, and 105 parameters, summarize the most important interactions in the full four-way table. (For additional comments on interpretation, see Davis, 1974; and Goodman, 1972a, 1972b.)

Evaluating the GGM

It is reasonable to judge a model selection algorithm by checking to see whether it finds a good model. This necessitates defining a good model. Two possible criteria come to mind here, one relating to R^2 and the other to chi-square. The best .80-adequate model could be defined as that model with the minimum number of parameters that yields $R^2 \geq .80$. Alternatively, the best model may be the one that yields an insignificant chi-square with the minimum number of parameters in the model. We have examined the performance of the GGM, relative to these two criteria, on the five-way table used throughout this article and on the 2 × 2 × 2 × 2 table from Goodman (1972b).

For the five-way table, the GGM selected the best $R^2 = .80$ model, and for the 2 × 2 × 2 × 2 table, it selected the best $R^2 = .99$ and $R^2 = .999$ models. Also, for the 2 × 2 × 2 × 2 table, the GGM selected the best model that used the chi-square criterion and significance level of .01. The calculations required to apply the GGM to the 2 × 2 × 2 × 2 table are presented in Tables 5 and 6. Comparing the analysis using the GGM with Goodman's analysis, we note that depending on the value of R^2_T, the GGM will select either the model chosen by Goodman or a more parsimonious model that explains almost as much variation.

We emphasize that these investigations of the GGM are empirical investigations, and the results are based on the analyses of only two contingency tables. However, on the basis of these analyses, the GGM does appear to perform well. We emphasize that there will generally be more than one minimal adequate model for a table. Additional insight into the structure of the table may be gained by locat-

Table 4
Results of Deleting Single β Parameters From Model H_6

β parameter removed	Resulting logit model	*df*	χ^2	*df* increase[a]	χ^2 increase[a]	R^2
RY	AS, AR, Y	770	1,938.51	84	845.48	.67
AS	AR, RY, S	689	1,680.93	3	587.90	.71
AR	AS, RY	728	1,479.77	42	386.74	.75

[a] Differences are measured with respect to corresponding statistics of Model H_6.

Table 5
Statistics for Pairs of Models to Select Entry Order of Effects in Models for the Data Analyzed in Goodman (1972b)

Abbreviated list of β parameters included in logit model	Fitted marginals	β parameter under consideration	*df*	χ^2	$C(\beta)$	Entry order
OR	(COR), (POR)	C	4	690.89	665.97	2
OR, C	(COR), (POR), (PC)		3	24.92		
RC	(COR), (PRC)	O	4	2,285.35	2,267.98	1
RC, O	(COR), (PRC), (PO)		3	17.37		
OC	(COR), (POC)	R	4	152.65	151.20	3
OC, R	(COR), (POC), (PR)		3	1.45		
OR, RC	(COR), (POR), (PRC)	CO	2	17.29	16.62	4
OR, RC, CO	(COR), (POR), (PRC), (PCO)		1	.67		
OR, OC	(COR), (POR), (POC)	CR	2	.68	.01	6
OR, OC, CR	(COR), (POR), (POC), (PCR)		1	.67		
OC, RC	(COR), (POC), (PRC)	OR	2	1.32	.65	5
OC, RC, OR	(COR), (POC), (PRC), (POR)		1	.67		

ing all minimal adequate models and comparing the alternative explanations offered by these models.

Discussion

The GGM is a flexible model selection procedure; it is not intended to be an automatic procedure into which the experimenter inserts a contingency table and an R_T^2 value and from which a model is received. Rather, it is proposed as a reasonable procedure for searching among the multitude of possible hierarchical models for appropriate models that satisfy specified criteria. If the researcher has little or no prior information, the procedure will help to identify reasonable models.

If, however, the researcher has additional information on the contingency table, the GGM can be modified to incorporate that information. For example, if it is known from previous studies that certain effects almost

Table 6
χ^2 and R^2 Statistics for a Nested Series of Models for the Table Analyzed in Goodman (1972b)

Model	Abbreviated list of β parameters included in logit model	Fitted marginals	β parameter added to previous logit model	*df*	χ^2	R^2
H_{13}	Grand *M*	(COR), (P)		7	3,111.47	
H_{14}	O	(COR), (PO)	O	6	801.60	.7424
H_{15}	O, C	(COR), (PO), (PC)	C	5	186.36	.9401
H_{16}	O, C, R	(COR), (PO), (PC), (PR)	R	4	24.96	.9920
H_{17}	CO, R	(COR), (PCO), (PR)	CO	3	1.45	.9995
H_{18}	CO, OR	(COR), (PCO), (POR)	OR	2	.68	.9998
H_{19}	CO, OR, CR	(COR), (PCO), (POR), (PCR)	CR	1	.67	.9998

Note. This series of models is derived from results in Table 5.

surely must be included in any model that is to fit the contingency table, then the model that includes those effects could be used as the base model for beginning the search, rather than Model H_0.

Another way that the GGM can be modified to reflect substantive knowledge that the researcher may have is that the choice among minimal adequate models need not necessarily center on the model with the fewest parameters. Given substantive considerations, it may be that one of the other minimal adequate models is far more interpretable or useful.

If the researcher prefers to use chi-square-based statistics rather than the R^2 statistics, the GGM can be easily modified to accomodate this. For example, the effects can be ordered by $\chi^2(H^*) - \chi^2(H^{**})$ rather than by $C(\beta)$.

The researcher may wish to restrict attention to models with no effect of order higher than some selected value, for example, three. If this is the case, then Step 1 in the GGM is easily modified to include an examination of only those effects that are candidates for inclusion in the model; no fourth or higher order effects would be examined in Step 1.

The GGM may also be based on an ordering of the effects by their contribution when each is entered as soon as possible into the model. In Step 1, instead of comparing H^* and H^{**} as defined earlier, the researcher would compare the Model H' that contains only the effect of interest and its lower order relatives, with the Model H'' containing all effects in H' except the effect of interest. For example, to obtain the contribution of AS, the Model H' [(ASRY), (AL), (SL), (ASL)] would be compared with H'' [(ASRY), (AL), (SL)]. We performed this procedure for the five-way table from Miller et al. (1974), which was used as an example in earlier sections, and for the table obtained from Goodman (1972b), which was analyzed in earlier sections. For the five-way table, the results were identical to those of the procedure in Step 1 of the GGM, which ordered the effects by their contribution when entered last into the model, except that the age and Age × Sex effects, which were close in value, were reversed. The two procedures yielded the same final model. In the analysis of the Goodman data, the two procedures yielded the same order of entry and the same final model. Thus the impact of a variable's order of entry on the ranking of its explanatory power appears to be much smaller here than is the case in multiple regression.

Only the class of nonhierarchical models is excluded from consideration by the GGM. Any hierarchical combination of main and higher order effects could be selected. Step 1 in the algorithm assures that all higher order effects will be examined. Thus important higher order interactions would not be missed, as they might be if we fit only models of uniform order, that is, models with all two-factor interactions, then models with all three-factor interactions, and so forth continuing until the lack of fit of a model of uniform order is insignificant.

The analyses described here are in the spirit of exploratory, rather than confirmatory, statistical analyses (Tukey, 1977); many tests are being performed on the same set of data after the fact. Hence, the significance probabilities obtained should be viewed as guidelines rather than as the results of formal tests.

In concluding, we advise, as do Bishop et al. (1975), that before a final decision on a model is made, the fit of the model should be examined cell by cell to check for outliers or any unusual patterns in the residuals.

Reference Note

1. Zahn, D. A. *Documentation for CONTAB: A computer program to aid in the analysis of multidimensional contingency tables using log-linear models* (Report No. M292). Tallahassee: Florida State University, Department of Statistics, March 1974.

References

Benedetti, J. K., & Brown, M. B. Strategies for the selection of log-linear models. *Biometrics*, 1978, *34*, 680–686.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. *Discrete multivariate analysis: Theory and practice.* Cambridge, Mass.: MIT Press, 1975.

Daniel, C., & Wood, F. S. *Fitting equations to data.* New York: Wiley Interscience, 1971.

Davis, J. A. Hierarchical models for significance tests in multivariate contingency tables: An exegesis of Goodman's recent papers. In H. L. Costner (Ed.), *Sociological methodology 1973–74.* San Francisco: Jossey-Bass, 1974.

Deming, W. E., & Stephan, F. F. On a least squares adjustment of a sampled frequency table when the

expected marginal totals are known. *Annals of Mathematical Statistics*, 1940, *11*, 427–444.

Dixon, W. J. (Ed.). *BMDP—Biomedical computer programs*. Berkeley: University of California Press, 1975.

Draper, N. H., & Smith, H. *Applied regression analysis*. New York: Wiley, 1966.

Fienberg, S. E. The analysis of multidimensional contingency tables. *Ecology*, 1970, *51*, 419–433.

Goldberger, A. S. Correlations between binary outcomes and probabilistic predictions. *Journal of the American Statistical Association*, 1973, *68*, 84.

Goodman, L. A. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, 1971, *13*, 33–61.

Goodman, L. A. A general model for the analysis of surveys. *American Journal of Sociology*, 1972, *77*, 1035–1086. (a)

Goodman, L. A. A modified multiple regression approach to the analysis of dichtomous variables. *American Sociological Review*, 1972, *37*, 28–46. (b)

Goodman, L. A. Guided and unguided methods for the selection of models for a set of *T* multidimensional contingency tables. *Journal of the American Statistical Association*, 1973, *68*, 165–175.

Grizzle, J. E., Starmer, C. F., & Koch, G. C. Analysis of categorical data by linear models. *Biometrics*, 1969, *25*, 489–504

Haberman, S. J. Loglinear fit for contingency tables (Algorithm AS 51). *Applied Statistics*, 1972, *21*, 218–225.

Higgins, J. E., & Koch, G. C. Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *International Statistical Review*, 1977, *45*, 51–62.

Hocking, R. R. The analysis and selection of variables in linear regression. *Biometrics*, 1976, *32*, 1–49.

Miller, K., Simons, R. L., & Fein, S. B. Compulsory mental hospitalization in England and Wales. *Journal of Health and Social Behavior*, 1974, *15*, 151–156.

Morrison, D. G. Upper bounds for correlations between binary outcomes and probabilistic predications. *Journal of the American Statistical Association*, 1972, *67*, 68–70.

Mosteller, F., & Tukey, J. W. *Data analysis and regression: A second course in statistics*. Reading, Mass.: Addison-Wesley, 1977.

Namboodiri, N. K., Carter, L. F., & Blalock, H. M., Jr. *Applied multivariate analysis and experimental designs*. New York: McGraw-Hill, 1975.

Shaffer, J. P. Defining and testing hypotheses in multidimensional contingency tables. *Psychological Bulletin*, 1973, *79*, 127–141.

Tukey, J. W. *Exploratory data analysis*. Reading, Mass.: Addison-Wesley, 1977.

Received January 23, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1201–1222

Superior–Subordinate Communication: The State of the Art

Fredric M. Jablin
Department of Speech Communication, University of Texas at Austin

Based on a review of the literature, empirical research in the area of superior–subordinate communication is classified into nine topical categories and critically examined. Inspection of this literature suggests that researchers have focused the majority of their attention on studying (a) the effects of power and status on superior–subordinate communication, (b) trust as a moderator of superior–subordinate communication, and (c) semantic-information distance as a source of misunderstanding in superior–subordinate communication. It is concluded that future research should increasingly be developmental and longitudinal in nature and should take into greater consideration the effects situational variables have on communication in the superior–subordinate dyad.

Status hierarchy is inherent in the nature of purposeful organizations. As Redding (1972) observes, within organizations "there are 'superiors' and 'subordinates'—even though these terms may not be expressly used, and even though there may exist fluid arrangements whereby superior and subordinates roles may be reversible" (p. 18). How superiors and subordinates interact and communicate to achieve both personal and organizational goals has been an object of investigation by social scientists for most of the 20th century. Empirical research examining superior–subordinate communication is diverse, is strewn across a multitude of disciplines, lacks coherent organization and classification, and in general, has not received sufficient review and interpretation as a body of literature. The present article attempts to alleviate this confusion by reviewing, classifying, interpreting, and providing directions for future research in the area of organizational communication that is loosely termed *superior–subordinate communication*.

This article focuses on empirical research solely in the domain of organizational communication. To avoid generalizations from communication research outside of the organizational environment, I do not review investigations exploring small group and interpersonal communication extraneous of purposeful organizations (with occasional exception). Since organizational communication is different, in a variety of ways, from communication in other settings (e.g., Redding, 1972, Rogers & Rogers, 1976) and given the difficulty of generalizing from social science research, regardless of area, limiting the setting (or scope) certainly adds to the validity of any knowledge claims. For example, it is difficult to generalize from small group communication research, which is external to organizational environments, to group communication within organizations, in which groups of groups are tied together in networks of networks. Hence, this review focuses on studies conducted within organizations or simulations of organizations.

In addition, this collection and critique of superior–subordinate communication research has excluded studies related to interviewing, despite the fact that they may have involved superior–subordinate interaction. (See Daly, Note 1, for a complete review of this literature.) Moreover, the nucleus of this review is the examination of interpersonal dyadic interactions between superiors and subordinates. Specifically, an attempt was made to avoid examination of research concerned with the use of impersonal, media-related (e.g., house organ, bulletin board, suggestion box)

Requests for reprints shoud be sent to Fredric M. Jablin, Department of Speech Communication, University of Texas at Austin, Austin, Texas 78712.

superior–subordinate communication. However, both written and oral face-to-face communication transactions, when of an interpersonal dyadic nature, were reviewed.

The article is organized into three sections; the first presents a basic definition of superior–subordinate communication. The second section reviews and organizes empirical research related to superior–subordinate communication into nine topical categories. The final section provides a discussion of the review and directions for future research.

Superior–Subordinate Communication Defined

The expressions superior and subordinate are derived from Latin roots, which when joined suggest that within an interpersonal relationship one individual is of subrank or is situated below another. In purposeful organizations, both formal and informal superior–subordinate relations usually exist. Moreover, most research evidence indicates that informal (i.e., not prescribed by organizational directives) superior–subordinate affiliations may be as important as formal veritable relations in determining communicative behavior. However, for the purposes of the present review, the definition of superior–subordinate communication is limited to those exchanges of information and influence between organizational members, at least one of whom has formal (as defined by official organizational sources) authority to direct and evaluate the activities of other organizational members.

Katz and Kahn (1966) provide probably the most parsimonious yet complete description of the types of communication that are typically exchanged in superior–subordinate interactions. These theorists suggest that downward communications from superior to subordinate are of five basic types: (a) job instructions, (b) job rationale, (c) organizational procedures and practices, (d) feedback about subordinate performance, and (e) indoctrination of goals (pp. 239–241). On the other hand, communication upward from subordinate to superior is reported to take four primary forms: (a) information about the subordinate himself/herself. (b) information about co-workers and their problems, (c) information about organizational practices and policies, and (d) information about what needs to be done and how it can be done (p. 245). More specific and detailed taxonomies of messages exchanged in superior–subordinate communication are available in the literature (e.g., Eilon, 1968; Melcher & Beller, 1967; Yoder, 1970).

Review of Literature

The empirical literature on superior–subordinate communication has been divided into nine topical categories.[1] Each of these categories represents a series of investigations that appear to be researching similar constructs from analogous theoretical foundations.[2] Many of the studies reviewed share more than one category but were classified into conceptually distinguishable groupings for purposes of clarification and parsimony.

Interaction Patterns and Related Attitudes

Researchers have investigated a variety of issues related to interaction patterns between superiors and subordinates. For example, numerous studies report that between one third and two thirds of a supervisor's time is spent in communicating with subordinates and that face-to-face discussion is the dominant mode of interaction (e.g., Berkowitz & Bennis, 1961; Brenner & Sigband, 1973; Dubin & Spray, 1964; Hinrichs, 1964; Kelly, 1964;

---

[1] The reader will note that this review contains no single category of research related to downward communication in superior–subordinate interaction. Since Redding (1972, see especially pp. 388–404) provides an extensive review of this literature prior to 1970 and given that less research has been pursued in this area subsequent to 1970, the present review has not directly focused attention on this area. Rather, research related to downward communication is discussed within the confines of the other categories.

[2] The reader will also note that this review does not consider research that could be classified as relating to participative decision making. Since exhaustive reviews of this literature already exist (e.g., Redding, 1972, pp. 154–250; Vroom, 1970, pp. 227–240; Vroom, 1976, pp. 1538–1546) further critique would be redundant.

Lawler, Porter, & Tenenbaum, 1968; Penfield, 1974). Moreover, results from a number of investigations indicate that the majority of superior–subordinate interaction concerns task issues (e.g., Baird, 1974; Richetto, 1969; Zima, 1969; Walton, Note 2) and that superiors and subordinates talk more about impersonal (focus of topics external to self) than about personal (directly related to self) topics (Baird, 1974). Further, research suggests that superiors are more likely to initiate interactions with subordinates than the other way around (e.g., Berkowitz & Bennis, 1961; Dubin & Spray, 1964). Yet, it is of interest to observe that superiors are less positive toward and less satisfied with interactions with their subordinates than they are with contacts with their bosses (e.g., Clement, 1974; Lawler et al., 1968; Tenenbaum, 1971). This finding is even more ironic when considered in light of Baird and Diebolt's (1976) discovery that a subordinate's job satisfaction is positively correlated with estimates of communication contact with superiors.

Several other studies present findings relevant to interaction patterns between superiors and subordinates. The results of these investigations suggest the following conclusions: (a) Superiors perceive that they communicate more with subordinates than subordinates perceive, whereas subordinates feel they send more messages to their superiors than the latter perceive (Webber, 1970); (b) superiors who lack self-confidence in their leadership abilities are less willing to hold face-to-face discussions with their subordinates than are superiors who are confident in their leadership abilities (Kipnis & Lane, 1962); (c) role conflict and role ambiguity are strongly correlated with "leader behavior indicative of direct as opposed to indirect interactions with subordinates" (Rizzo, House, & Lirtzmann, 1970, p. 162); (d) when a subordinate needs informal help in the work setting, he/she is more likely to seek assistance from his/her superior than peers or subordinates (as reported by helpees) (Burke, Weir, & Duncan, 1976); and (e) supervisors are more likely to serve as "production" communication liaisons than as "maintenance" or "innovation" liaisons (MacDonald, 1976).

In summary, studies that explored interaction patterns between superiors and subordinates suggest frequent task-oriented communication within the dyad but differential attitudes and perceptions of those interactions. Moreover, personal characteristics and needs of the interactants seem to mediate their desire for and perceptions of superior–subordinate communication.

Openness in Communication

Two basic dimensions of openness in superior–subordinate communication can be distinguished: openness in message sending and openness in message receiving. Redding (1972) describes openness in message sending as the "candid disclosure of feelings, or 'bad news,' and important company facts" (p. 330), whereas openness in message receiving involves "encouraging, or at least permitting, the frank expression of views divergent from one's own; the willingness to listen to 'bad news' or discomforting information" (p. 330). Baird (1974) adds that it is essential that researchers clearly specify whether they are referring to task-relevant openness or non-task-relevant openness when investigating each of the above dimensions.

Much of the impetus for studying openness in superior–subordinate communication has been provided by management theorists who have suggested that openness is an essential element for an effective organizational climate (e.g., Haney, 1967; Likert, 1967). Support for this proposition is furnished in studies by Burke and Wilcox (1969), Baird (1974), and Jablin (1978a), who have found that employees are more satisfied with their jobs when openness of communication exists between subordinate and superior. Furthermore, several inquiries report that openness of communication is directly correlated with organizational performance (e.g., Indik, Georgopoulos, & Seashore, 1961; Willits, 1967). However, it should be noted that the results of one investigation suggest that managerial effectiveness is unrelated to openness of communication between superior and subordinate (Rubin & Goldman, 1968).

A series of doctoral dissertations completed at Purdue University have attempted to ex-

plore in detail the communication characteristics of openness in superior–subordinate relationships. The first of these researches (Baird, 1974) examined subordinates' "upward communication freedom" with superiors. Results of the study revealed that willingness of superiors and subordinates to talk as well as actual talk about a topic is a function of each interactant's perception of the other's willingness to listen. Extrapolating on Baird's study, Stull (1975) investigated superior and subordinate attitudes toward various types of supervisory responses to task-relevant and non-task-relevant open messages sent by subordinates. Analyses disclosed that for task and nontask topics, subordinates and superiors preferred supervisory responses that were accepting (encouraging) or reciprocating ("owning-up" to one's feelings, ideas, etc.) rather than neutral-negative (unfeeling, cold, or nonaccepting). Finally Jablin (1978a, 1978b), attempting to determine the types of communicative responses that characterize open and closed relationships between superiors and subordinates, experimentally studied the attitudes of subordinates toward five basic types of message responses occurring in a dyad: confirmation (a response that provides a speaker with positive content and positive relational feedback), disagreement (a response that provides a speaker with negative content feedback but positive relational feedback), accedence (a response that provides a speaker with positive content feedback but negative relational feedback), repudiation (a response that provides a speaker with both negative content and negative relational feedback), and disconfirmation (a response that provides a speaker with irrelevant content and equally irrelevant relational feedback). Results from the investigation indicated (a) that disconfirming responses are not acceptable in superior–subordinate communication; (b) that subordinates prefer message responses from superiors that provide positive relational feedback; (c) that regardless of perceived openness or closedness of the communication relationship with their superior, subordinates expected the same types of responses from a superior but evaluated the appropriateness of these responses differently; (d) that a substantial degree of reciprocity exists for confirming messages, regardless of the openness or closedness of the superior–subordinate relationship; and (e) that subordinates who perceive a closed relationship with their superior are prepared to respond to a superiors' message, which contains negative relational feedback toward the subordinate, with a response transmitting negative relational feedback toward the superior; however, this is not true for subordinates who perceive an open relationship with their superior.

In summary, these studies suggest that in an open communication relationship between superior and subordinate, both parties perceive the other interactant as a willing and receptive listener and refrain from responses that might be perceived as providing negative relational or disconfirming feedback. Moreover, these inquiries suggest that what distinguishes an open from a closed superior–subordinate relationship may not be the types of messages exchanged but how the interactants evaluate the appropriateness of these communications. Finally, these studies provide strong evidence for the proposition that employees are more satisfied with their jobs when openness of communication exists between superior and subordinate than when the relationship is closed.

Upward Distortion

Closely related to research examining openness in superior–subordinate relationships are a group of investigations exploring message distortion in subordinate upward communication to superiors. Mellinger (1956), who collected questionnaire data from 330 scientists in a medical laboratory, is generally credited with the initiation of this research tradition. (For a discussion of related research antecedent to Mellinger's investigation, see Guetzkow, 1965, pp. 553–555.) Results of this early inquiry into message distortion revealed that when Individual A does not trust Individual B, Individual A will conceal his/her feelings when communicating to B about a particular issue. Moreover, concealment of Individual A's true feelings was found to be often associated with evasive, compliant, or aggressive communicative behavior on his/her

part and with under- or overestimation of agreement on the issue by Individual B. Cohen's (1958) replication and clarification of Kelly's (1951) investigation of upward communication in experimentally created hierarchies also inspired a tradition of research in the area of upward communication distortion. Results of this study suggested that within a hierarchy, if an individual has power over the advancement of persons of lower rank, those of lower rank will omit critical comments in their communication with the person of higher rank. Thus with these seminal studies, Mellinger (1956) and Cohen (1958) initiated a sphere of research that examined the moderating effects of trust and mobility aspirations on upward communication distortion.

Building on the previously described research, Read and his associates (Maier, Hoffman, & Read, 1963; Read, 1962) explored the relationships among upward mobility aspirations, trust, and the accuracy with which managers communicate information upward in organizational hierarchies. Data analyses supported the earlier findings of Mellinger and Cohen, which indicated that mobility aspirations (i.e., desire for advancement and status seeking proclivity) and low trust in one's superior are negatively related to accuracy of upward communication. Moreover, results suggested that even when a subordinate trusts his/her superior, high mobility aspirations "strongly militate against accurate communication of potentially threatening information" (Read, 1962, p. 13). In addition, it was discovered that "subordinates feel less free to communicate with superiors who previously have held their position than with those who have not" (Maier et al., 1963, p. 9).

More recently, research by Roberts and O'Reilly (1974) and O'Reilly and Roberts (1974) has supported the notion that a subordinate's trust in his/her superior is a facilitator of distortion-free upward communication. However, these researchers did not find strong correlations between subject's mobility aspirations and propensity towards upward communication distortion. Finally Sussman (1974), investigating upward communication distortion from the perspective of the superior (i.e., the recipient of distorted messages), failed to find that superiors perceive greater accuracy in messages from subordinates who are perceived as trusting the superior than in messages from subordinates perceived as nontrusting.

In contrast to these studies, several researchers have investigated the origins and concomitants of upward communication distortion from slightly different perspectives. Athanassiades (1973, 1974) argues that ascendency and security needs, risk-taking propensity, and organizational climate, when perceived as instrumental to a subordinate's goals, will produce upward communication distortion. Results of his research indicate that for both male and female subordinates, upward distortion is need motivated, with distortion being positively related to achievement needs and negatively related to level of security. Furthermore, Athanassiades's findings suggest that distortion of upward communication is negatively related to an autonomous organizational climate and positively related to a heteronomous climate. It is of interest to observe that his findings (1974) also show that "women in managerial positions feel more suppressed—less autonomous, less independent—than men do in similar positions" (p. 208).

A recent study by Young (1978) supports Athanassiades's finding that organizational climate is related to distortion of upward communication. Specifically, results suggest that in organic as compared with mechanistic organizational environments, subordinates perceive greater appropriateness, expect fewer harmful consequences, and evidence greater willingness to disclose important yet personally threatening information to superiors. However, his data also disclose that the upward communication behavior of female subordinates follows more closely the behavior of subordinates in an organic work setting than does the upward communication of male subordinates.

Extrapolating on Athanassiades's research, Level and Johnson (1978) have recently found that upward distortion is most likely to be associated with messages in which in-

formation about the following personality factors is communicated: ascendency, responsibility, emotional stability, cautiousness, and original thinking. Their data also suggest that in certain areas subordinate tendencies to distort upward communication can be reduced by increasing the superior's "consideration" leadership style, or increasing the accuracy with which the superior transmits downward information. In addition, Krivonos (1976) examined the role of motivation theory in upward communication distortion and found that superiors perceive that intrinsically motivated subordinates distort messages less than do extrinsically motivated subordinates.

Research that investigates types of messages that tend to be distorted in upward communication indicates that subordinates will be less reluctant to communicate information that is positive–favorable than negative–unfavorable (O'Reilly & Roberts, 1974; Rosen & Tesser, 1970) and that superiors view messages that are favorable to subordinates as less accurate than messages that are unfavorable to subordinates (Sussman, 1974). Moreover, O'Reilly and Roberts (1974) report that when information is both favorable and important, subordinates do not hesitate to communicate it upward to their superiors. In addition, Housel and Davis (1977) have discovered that subordinate satisfaction with upward communication tends to vary as a function of the channel used; face-to-face channels are most satisfactory, followed by telephone and written channels. Finally, a recent study of Rosen and Adams (1974), which examined the severity of discipline administered to subordinates who distorted upward communication, reveals that "recommended disciplinary measures were relatively mild when the subordinate's motives were altruistic and when his superior was dependent on him for expertise" (p. 382).

In summary, research that explored upward distortion in superior–subordinate communication has examined numerous variables that may moderate the occurrence of upward distortion. These variables include trust, mobility aspirations, ascendency and security needs, organizational climate, sex differences, motivation, message characteristics, and a variety of upward communication channels. Although the effects of subordinate trust in superior, subordinate mobility aspirations/ascendency needs, and the contingent role of organizational climate on upward distortion seem best supported, evaluation of the research in total suggests that probably no one variable can sufficiently explain the phenomena and that additional multivariate research is required before we can place confidence in any one of these explanations.

Upward Influence

Influence processes are a central feature of superior–subordinate communication. And, as Walter (1966) notes,

> to study influence, one must first study communication, for influence without communication is as wildly implausible as action at a distance. Influence is always accompanied by some form of communication, blunt or subtle, overt or tacit: Advertising, lobbying, arguing a case before a jury or on a suitor's knee. (p. 190)

In studies of superior–subordinate communication, researchers have focused attention on two basic dimensions of influence: (a) the effects a superior's influence in the organizational hierarchy has on his/her relationships with subordinates and (b) the transmission of influence by subordinates to superiors. Research in this latter category is diffused and varied, and since it is represented in other sections of this article (e.g., upward distortion, feedback), it is not directly discussed here.

Due to its effects on superior–subordinate relations, the upward influence of a subordinate's superior with his/her boss has received considerable attention in the research literature. Probably best known is the so-called "Pelz effect." In his seminal study Pelz (1952), who collected data from over 8,000 supervisory and nonsupervisory personnel in the Detriot Edison Company, discovered that a superior's upward influence in an organization moderates his/her social closeness with subordinates. Specifically, he found that "employee-centered supervisors are associated with higher levels of employee satisfaction *only when the supervisor apparently exercises influence 'upward' with his own superiors*"

(Redding, 1972, p. 438). More recently Wager (1965), who explored leadership behaviors and influence in one organization, reported findings similar to those of Pelz but also observed that the magnitude of the moderating effect of influence varied positively with the organizational status of the respondent. At this point it is important to note that Pelz's and Wager's influence measures were concerned only with supervisory influence with respect to personnel management of subordinates and did not assess influence in such areas as resource allocation, organizational changes, policy formation, or objective setting (Wager, 1965). Moreover, this typifies most of the research in this area.

In recent years investigators have once again begun to explore the relationship between superior's upward influence and communication with subordinates. For example, House, Filley, and Gujarati (1971) report that the interaction between superior's hierarchical influence and consideration behavior with subordinates varies from company to company (i.e., it is situational). Perhaps of even greater importance is their finding that when a superior is too high in upward influence, dysfunctional consequences may emerge in relation to subordinate willingness to openly communicate with the superior. They argue that

> where supervisors are seen to have such high influence, it is likely that there will be greater status separation between them and their subordinates, and that such status differentiation will result in a restriction of upward information flow, less willingness on the part of subordinates to approach superiors, and less satisfaction with the social climate of the work unit. (p. 429)

In a related study, Roberts and O'Reilly (1974) report that subordinates who perceive their superior as having high upward influence also have a high desire for interaction with the superior, high trust in the superior, and a high estimation of accuracy of information received from the superior. Similarly, Jones, James, and Bruni (1975) have found that subordinate confidence and trust in a superior is positively related to the superior's success in interactions with higher levels of management. Finally, O'Reilly and Roberts (1974) and Roberts and O'Reilly (1974), who examined the association between superior upward influence and subordinate upward communication distortion, have discovered only weak correlations between these variables.

In summary, results from studies that explored the relationship between superior's upward influence and communication with subordinates suggest the following conclusion: Subordinates who perceive their superior as having substantial but not excessive upward influence with their bosses will be more satisfied with their superior and will interact and trust him/her more than will subordinates who perceive their superior as low in upward influence. However, since several of the previously described investigations indicate that this conclusion may be situation bound and may be contingent on factors other than those already studied, the conclusion warrants only tentative acceptance.

Semantic-Information Distance

Originally coined by Tompkins (1962), the term *semantic-information distance* describes the gap in information and understanding that exists between superiors and subordinates (or other groups within an organization) on specified issues. This concept is analogous to the concept of "disparity" advanced by Browne and Neitzel (1952), to Weaver's (1958) construct of "semantic barrier," to "categorical and syndectic similarity" as proposed by Triandis (1959a, 1959b, 1959c, 1960), to "semantic agreement" as discussed in Maier, Hoffman, Hooven, and Read (1961) and Maier et al. (1963), and to "congruence" as explored in research by Minter (1969). Studies that examined the nature and definitional qualities of semantic-information distance through 1970 are discussed in detail in Redding's (1972) review of organizational communication literature.

The basic conclusions that can be drawn from the early research on semantic-information distance can be briefly described as follows: (a) The larger the semantic distance between superior and subordinate, the lower will be the subordinate's morale (Browne & Neitzel, 1952); (b) superiors tend to overestimate the amount of knowledge subordi-

nate's possess on given topics (Odiorne, 1954); (c) management personnel tend to describe themselves by traits that are different from those that subordinates use to describe themselves (Porter, 1958); (d) managers and workers differ in the criteria that they use in making judgments about people (Triandis, 1959a, 1959b, 1959c, 1960); (e) significant gaps in semantic distance exist between union and management (Schwartz, Stark, & Schiffman, 1970; Weaver, 1958) and between union leadership and their members (Tompkins, 1962); (f) superiors and subordinates have difficulty agreeing on the basic job duties and demands facing subordinates (Maier et al., 1961; Rosen, 1961); (g) whether a superior has previously held his/her subordinate's job has little effect on reducing the semantic-information distance between them (Maier et al., 1963); (h) superior's perceptions of the attitudes of subordinates toward him/her is often unrelated to their actual attitudes (Bowers, 1963; White, 1976); (i) serious semantic differences between superior and subordinate are frequent (e.g., Minter, 1969, reports that they occur over 60% of the time); and (j) there is some evidence that indicates that superiors "find it easier to communicate with subordinate managers whose attitudes are similar [rather than dissimilar] to their own" (Miles, 1964, p. 324).

In general, most research related to semantic-information distance conducted since 1970 is supportive of the aforementioned studies. Greene (1972) has found that the more accurately a subordinate complies with his/her superior's expectations of subordinate behavior, the higher the subordinate's job satisfaction and the better his/her performance evaluation by the superior. Supportive of the results of Greene's research is a study by Pfeffer and Salancik (1975) that suggests that the behavior of subordinates and superiors is constrained by the expectations of other members of the role set. Several recent investigations contribute to the list of areas in which significant superior–subordinate semantic-information distance exists. Examining superior–subordinate dyads, Boyd and Jensen (1972) found that first-line managers and their superiors experience difficulty in agreeing on the authority of the first-line manager, whereas Moore (1974) reports that a new manager's superior and his/her subordinates tend to disagree on how long it will take the manager to learn the new position. Assuming that empathic ability is negatively correlated to semantic-information distance, Northouse's (1977) study would strongly indicate that one means of reducing semantic distance is by increasing trust between superior and subordinate. Finally, a study by Baird and Diebolt (1976) found no relationships between superior–subordinate role congruence and several communication variables; however, limitations within the study restrict the generalizability of its findings.

In summary, results of empirical research in the area of superior–subordinate semantic-information distance probably provide some of the most consistent conclusions of any topic of study in organizational communication. Incessantly, we find the existence of semantic-information distance in superior–subordinate relations, often at levels that would appear to seriously obstruct organizational effectiveness. The catalogue of topical areas in which semantic differences between superiors and subordinates tend to occur is expanding and would strongly suggest that future research should pursue the development of valid and reliable techniques to reduce this semantic gap.

Effective Versus Ineffective Superiors

Interest in identifying the communicative behaviors of effective leaders probably has existed since the earliest days of civilization, when humankind became proficient at organizing for battlefield warfare and thus required an expendable supply of effective leaders. Hence, over the years the identification of effective as compared to ineffective communication behaviors of superiors has received more investigation than any other area of organizational communication.

From the period of 1950 to the mid-1960s, a series of doctoral dissertations completed at Purdue University attempted to determine the communication correlates of "good" supervisors (Funk, 1956; Kelly, 1963; Minter,

1969; Miraglia, 1964; Pyron, 1964; Richetto, 1969; Simons, 1962; Sincoff, 1970; Smith, 1968; Zima, 1969). For the majority of these studies, good supervision as compared to poor supervision was determined by higher management evaluation of supervisors. Redding (1972, pp. 436–446) succinctly summarizes the results of these researchers and suggests the following general conclusions:

> 1. The better supervisors tend to be more "communication-minded"; e.g., they enjoy talking and speaking up in meetings; they are able to explain instructions and policies; they enjoy conversing with subordinates. (See especially Funk, 1956; Pyron, 1964.)
>
> 2. The better supervisors tend to be willing, empathic listeners; they respond understandingly to so-called "silly" questions from employees; they are approachable; they will listen to suggestions and complaints, with an attitude of fair consideration and willingness to take appropriate action. (See especially Funk, 1956; Simons, 1962; Kelly, 1963; Zima, 1969.)
>
> 3. The better supervisors tend (with some notable exceptions) to "ask" or "persuade," in preference to "telling" or "demanding." (See especially Simons, 1962; Pyron, 1964.)
>
> 4. The better supervisors tend to be sensitive to the feelings and ego-defense needs of their subordinates; e.g., they are careful to reprimand in private rather than in public. (See, e.g., Simons, 1962.)
>
> 5. The better supervisors tend to be more open in their passing along of information; they are in favor of giving advance notice of impending changes, and of explaining the "reasons why" behind policies and regulations. (See especially Funk, 1956; Simons, 1962.) (Redding, 1972, p. 443)

Other research on superior–subordinate communication contemporary to that of the Purdue group generally supports the thrust of the above conclusions (Brown, 1964; Jain, 1971; Ponder, 1959; Sadler, 1970; Tacey, 1959, Walker, Turner, & Guest, 1956).

For example, Ponder (1959) reports that effective as compared to ineffective foremen tend to be better communicators: They spend more time with employees carrying out the job, providing general supervision, and handling personnel matters. Moreover, more recent research provides additional testimony to the validity of the claims of these investigators (Duffy, 1975; Heizer, 1972; Sank, 1974; White, 1972) or has attempted to further elucidate the communication behaviors associated with various managerial interaction styles (e.g., Bradley & Baird, 1977).

Despite the strong evidence that characterizes the communication profile of effective superiors, other research suggests that effective supervisory communication behaviors are situational and contingent on a variety of factors (e.g., Downs & Pickett, 1977). As Redding (1972) notes in his own review of the Purdue studies, "The precise combination of behaviors or attitudes which 'works' in one company is likely to be different from what 'works' in another company or organization" (p. 445). The importance of viewing effective as compared to ineffective superior communication behavior from a contingency perspective is demonstrated by developments in three areas of leadership research: (a) the traditional study of leadership that employs the "consideration-initiating structure" framework (e.g., Fleishman & Harris, 1962), (b) Fiedler's contingency approach to leadership (e.g., Fiedler, 1967), and (c) a more recent view of leadership that uses a dyadic linkage – role-making model (e.g., Dansereau, Graen, & Haga, 1975).

As a result of extensive leadership research conducted at Ohio State University during the 1950s and early 1960s, two basic dimensions of leadership behavior were identified: (a) "consideration" and (b) "initiating structure." (See Stogdill, 1974, pp. 128–141, for a complete review of these studies.) Leader consideration was found to be typified by friendship and warmth, mutual trust, rapport and tolerance, and two-way communication between a leader and his/her work group (Fleishman, Harris, & Burtt, 1955). Initiating structure includes "behaviors in which the supervisor organizes and redefines group activities and his relation to the group . . . . This dimension seems to emphasize overt attempts to achieve organizational goals" (Fleishman & Harris, 1962, p. 43). These two basic dimensions of leader behavior are analogous to those denoted as "employee orientation" and "production orientation" in the University of Michigan Institute for Social Research leadership studies (e.g., Katz, Maccoby, & Morse, 1950). The importance of the above investigations to the identification of the effective communication behaviors of supervisors rests on the similarity between the constructs of consideration and employee

orientation, and communication. For example, Miraglia (1964), who studied the parallels between consideration and communication ability, discovered that consideration "is largely a matter of *communication* behavior" (Redding, 1972, p. 148). This conclusion has been supported in research by Jain (1973) and more recently in a study by Dennis (1974).

The general conclusion drawn from the early consideration and initiating structure research was that superiors are "rated as more effective when they score high in both consideration and leadership structure" (Stogdill, 1974, p. 140). Perhaps of even greater significance for communication researchers is the general finding that leaders high in consideration (good communicators) can increase structure within their work groups and still be rated as effective leaders (e.g., Fleishman & Harris, 1962). However, current inquiries suggest that numerous situational variables impinge on the validity and reliability of consideration-initiating structure (C-IS) to predict leader effectiveness. Reviewing this literature through 1973, Kerr, Schriesheim, Murphy, and Stogdill (1974) identify the following as situational variables that moderate the C-IS ability to predict leader behavior and performance:

> subordinate need for information, job level, subordinate expectations of leader behavior, perceived organizational independence, leader's similarity of attitudes and behavior to managerial style of management, leader upward influence; and characteristics of the task, including pressure and provision of intrinsic satisfaction. (p. 62)

More recent investigations also suggest that the following situational variables may moderate the C-IS predictive capability: sex (Day & Stogdill, 1972), task type (Hill & Hughs, 1974), length of employment and organizational climate (Kavanagh, 1975), and work-unit size (Schriesheim & Murphy, 1976).

Fiedler's work on leadership likewise indicates that researchers should be examining the communication attributes of effective superiors from a contingency perspective (e.g., Fiedler, 1964, 1967, 1970, 1971a, 1971b, 1972a, 1972b). Emphasizing the role of leader personality and style of interaction on work-group performance, Fiedler argues that three dimensions of task *situations* primarily determine leader effectiveness: leader–member relations (which obviously are dominated by communication behavior), task structure and leader-position power. In essence, this research suggests that supervisors have predominant styles of interacting with subordinates and that their effectiveness will vary, depending on whether the situation (as previously defined) is best suited to that style.

Finally, recent research that views leadership from a dyadic linkage – role-making perspective shows that supervisors do not develop the same kinds of relationships with all subordinates and superiors, and thus the communicative behavior that may be effective in one type of relationship may not be effective in another (Cashman, Dansereau, Graen, & Haga, 1976; Dansereau, Cashman, & Graen, 1973; Dansereau et al., 1975; Graen, Cashman, Ginsburgh, & Schiemann, 1977; Haga, Graen, & Dansereau, 1974). Specifically, results of this research indicate that supervisors tend to develop one of two types of exchange patterns with subordinates, that is, either a pattern of leadership exchange (characterized by "influence over a member without resort to authority" [Cashman, 1976, p. 281]) or supervision exchange (in which "influence over a member is based primarily upon authority" [Cashman et al., 1976, p. 281]). Moreover, it has been found that supervisors in turn develop either leadership exchanges or supervision exchanges with their bosses. In addition, findings suggest that

> subordinate members of the upper dyad who develop leadership exchanges with their bosses have greater influence with their bosses and receive more latitude, support and attention from their bosses than their colleagues who fail to develop leadership exchanges. (Graen et al., 1977, p. 502)

In summary, investigations that explored a dyadic linkage – role-making model of leadership suggest that superior–subordinate communication patterns are not stable across all superior–subordinate interactions and may vary as a function of organizational understructure.

As noted in the opening of this section, we are rich in research studies that have attempted to identify the effective and ineffec-

tive communication correlates of supervision. Clear evidence has been presented that suggests a certain profile that characterizes the communication behaviors of effective supervisors. On the other hand, several other research traditions were reviewed which indicate that the qualities of effective leadership vary from situation to situation and are contingent on numerous factors. Which set of findings and conclusions are we to believe? The answer would appear to be both. Data that comprise the Purdue and other related effectiveness studies have been collected from a myriad of organizations and supervision situations—the pattern of results is too consistent to reject without further research. It may be that for 60% of the superior–subordinate communication situations, these findings are applicable, yet they may only apply to 10% of the cases. Obviously, the only way we will be able to resolve this question is by research that investigates the effects situational variables have on superior–subordinate communication.

Personal Characteristics

In the process of studying superior–subordinate communication, researchers have attempted to discover the personal characteristics of members of that dyad that mediate their communication behavior. Since the variables examined in these investigations are extremely diverse and at present lack coherent organization, this section endeavors to provide such structure.

Several investigators have explored the effects interactant tendencies for internal as compared to external locus of control have on superior–subordinate interaction (Durand & Nord, 1976; Mitchell, Smyser, & Weed, 1975). Findings from these studies suggest (a) that internal subordinates see their supervisors as more considerate than do externals; (b) that internals are most satisfied with participative superiors, whereas externals are most satisfied with directive superiors; and (c) that internal superiors tend to use persuasion to obtain subordinate cooperation, whereas externals rely more on coercive power. In addition, research related to the study of locus of control indicates that subordinates and superiors with passive personalities tend to exaggerate the volume of their interaction with others, whereas active persons tend to underestimate their interaction (Webber, 1970).

Studies that examined the characteristics of supervisors and their communication-related behavior with subordinates suggest the following interesting conclusions: (a) High-least-preferred-co-worker (LPC) leaders under conditions of threat tend to engage in considerate behavior, whereas low-LPC leaders tend to increase initiating structure (Green, Nebeker, & Boni, 1976); (b) lower level supervisors tend to be more dogmatic than upper-middle and top-level managers (Close, 1975); (c) young managers (20–29 years) tend to be more autocratic and low in human relations skills than are middle-aged (30–40 years) or late middle-aged (40–55 years) managers (Pinder & Pinto, 1974); and (d) superiors tend to rate subordinates as competent when they have values similar to those of the superior (Senger, 1971).

On the other hand, studies exploring subordinates' perceptions of superior's communication behavior and personality indicate (a) that superiors who are apprehensive communicators are not particularly liked by subordinates (Daly, McCroskey, & Falcione, Note 3); (b) that subordinate's satisfaction with superiors can be predicted from several dimensions of homophily–heterophily (Daly, McCroskey, & Falcione, Note 4); (c) that authoritarian subordinates are most satisfied when they work for directive superiors (Bass, Valenzi, Farrow, & Solomon, 1975; Tosi, 1973); (d) that subordinate satisfaction with immediate supervision is related to subordinate perception of superior's credibility (Falcione, 1974); (e) that confirmation of subordinate's needs for affection and dominance results in greater perceived frequency of interaction between superior and subordinate (Hawkins, 1976); (f) that subordinates in small work groups, who require high interaction with co-workers and superiors and high interdependence, have negative attitudes toward authoritarian supervisors, whereas subordinates in large work groups, with restricted interaction and highly independent work, have more positive attitudes toward

authoritarian supervision (Vroom & Mann, 1960); and (g) that subordinates, regardless of their personality, tend to be most satisfied with superiors high in human relations orientation (Weed, Mitchell, & Moffitt, 1976). In addition, it should be noted that Hall (1974, 1975) has conducted a series of investigations that examine, in part, personality correlates of organizational members that affect superior–subordinate communication, the results of which are too voluminous to report here.

An area of superior–subordinate research that has received considerable attention of late is concerned with differences between male and female supervisory behaviors. The general results of these inquiries indicate that subordinates do not describe the behaviors of male and female leaders differently (e.g., Bartol & Wortman, 1975; Day & Stogdill, 1972) but do agree on the existence of leader sex role stereotypes (e.g., Rosen & Jerdee, 1973; Schein, 1973, 1975). Specifically, there is strong evidence which suggests that subordinates of both sexes are more satisfied with consideration behavior from a female superior than a male and are more satisfied with exhibition of initiating structure by male superiors than with similar behaviors from female superiors (Bartol & Butterfield, 1976; Petty & Lee, 1975; Petty & Miles, 1976). Moreover, Sussman, Pickett, Berzinski, and Pearce (in press) report that the sexual composition of superior–subordinate dyads does impose "norms and restrictions" on upward communication in the dyad.

In summary, studies that examined the effects of personal characteristics on superior–subordinate communication have tended to focus on three basic areas: (a) single characteristics of superiors or subordinates that affect their communication behavior, (b) characteristics of superiors and subordinates, taken together, and their effect on superior–subordinate communication, and (c) differences in superior–subordinate communication as a function of the sex of the interactants. Although the findings of most of this research are interesting and important, on the whole they tend to be isolated and to lack theoretical foundations. Future investigations should endeavor to remedy this situation by relating such studies to the larger scope of organizational communication theory.

Feedback

Probably one of the most common complaints aired by superiors and subordinates about their communication relationship is that one of the interactants does not provide the other with sufficient and relevant feedback. Both upward and downward feedback appear to be essential for effective superior–subordinate relations, since such feedback provides information that denotes the success or failure of policies and objectives, that suggests the need for corrective actions and controlling mechanisms, and that provides the members of the dyad with knowledge of the other party's sentiments about formal and informal organizational activities. In his collection of theory and research in organizational communication, Redding (1972) provides an extensive review of empirical inquiries that explore feedback in superior–subordinate communication through about 1970 (see especially pp. 39–62). Now classic studies such as Leavitt and Mueller (1951), Smith and Kight (1959), Gibb (1961), Zajonc (1962), Bowman (1963), Haney (1964), Meyer, Kay, and French (1965), Cook (1968), and Minter (1969) are summarized in Redding's anthology. Hence, the following review examines only empirical research conducted in the area of feedback in superior–subordinate communication after 1970.

Studies that examined feedback in superior–subordinate communication since 1970 can be grouped into one of two categories: (a) investigations that explored the effects of subordinate's feedback to superiors on superior's behavior and (b) research analyzing the effects of superior's feedback to subordinates on subordinate's behavior. Investigations in the former category report a number of interesting findings. Brenner and Sigband (1973), who surveyed over 700 managers in a major aerospace firm have found that subordinate's feedback to superiors is greater when

(a) subordinates were told what was to be done with completed assignments, (b) the superior

formerly held the subordinate's position, (c) the superior made the largest proportion of assignments to the subordinate, and (d) the subordinate felt that he could secure clarification of assignments from his immediate superior. (p. 325)

Attempting to determine if a leader's verbal behavior could be altered by manipulating feedback to him/her, Butler and Jaffee (1974) report that positive feedback to a leader made him/her more task oriented, whereas negative feedback increased negative social-emotional behavior (as classified by Bales's Interaction Process Analysis category system). These researchers argue that their results

> indicate that in a production-oriented organization, positive feedback is to be preferred to negative feedback, and negative feedback might have very little to offer if no specific suggestions for changing one's behavior are given. (p. 335)

In a related study, Fodor (1974) explored the effects of a subordinate's disparagement of a superior's competence on the superior's distributions of rewards to subordinates. Results indicated that the superior tended to favor a compliant subordinate who was not an ingratiator. Finally, several studies have also attempted to determine whether the subordinate's feedback to superiors elicits changes in the superior's behavior and, concomitantly, changes in the subordinate's attitudes toward the superior. For example, Hegarty (1974), who used survey feedback methods, found that supervisory performance improved subsequent to subordinate feedback, whereas Burnaska (1976) relates findings which suggest that feedback and subsequent training can quickly change a supervisor's behavior but that worker perceptions of the superior change only with time.

A number of investigations have examined the effects of a superior's feedback to subordinates on the behavior of those subordinates. Harvey and Boettger (1971), in an experiment designed to improve communication in a managerial work group, reported a norm among subordinates against asking superiors for clarification of memos that are unclear or contain double messages. In a preliminary investigation that explored sources of feedback, Greller and Herold (1975) suggest that intrinsic (i.e., psychologically close to the individual) sources of information are seen by workers as providing more feedback than sources that are seen as external (i.e., psychologically distant). Moreover, in a related study, Kim and Hamner (1976) provide evidence that evaluative supervisory feedback to subordinate performance (i.e., extrinsic feedback) and nonevaluative feedback (i.e., subordinate self-generated or intrinsic feedback), when combined with a goal-setting production program, increase subordinate performance significantly beyond that of groups involved in just goal setting. Recent research also suggests (a) that superiors' feedback to a subordinate, which shows a lack of trust in the subordinate, results in subordinate dissatisfaction and aggressive feelings (Brenenstuhl, 1976); (b) that superiors perceived as expressive (high in human relations) are more likely to provide subordinates with social approval than those superiors perceived as instrumental (a Weberian orientation) (Marcus & House, 1973); (c) that in conflict situations supervisory responses that relate acceptance and encouragement of subordinate disagreement are associated with high subordinate satisfaction (Burke, 1970; Renwick, 1975); and (d) that under low surveillance (infrequent need to report to superior) positive feedback from superior to subordinate leads to greater subordinate compliance than when the subordinate receives no direct feedback, whereas under high surveillance conditions subordinates who receive positive feedback from their superiors comply less than when they receive no direct feedback from their superiors (Organ, 1974).

In addition, a variety of research has explored the effects of superior reward or reinforcement behavior on subordinate performance and satisfaction. These studies suggest the following conclusions: (a) Leader positive reward behaviors (e.g., recognition of subordinate performance) are generally associated with subordinate satisfaction, but the relationship between leader punitive rewards (e.g., corrective actions) and subordinate satisfaction varies as a function of the nature of the task performed by each work group (Sims & Szilagyi, 1975); (b) superiors who frequently criticize their subordinates for

poor work are generally rated as less effective than those who criticize less frequently (Oldham, 1976); and (c) a superior tends to positively reinforce a subordinate when he/she is positively reinforced by the subordinate's performance and to negatively reinforce a subordinate when he/she is negatively reinforced as a result of the subordinate's performance (Barrow, 1976; Greene, 1975; Hinton & Barrow, 1975; Lowin & Craig, 1968).

It should also be noted that research discussed earlier by Stull (1975) and Jablin (1978a) indicates that both superiors and subordinates prefer message responses from one another that provide positive relational feedback to the source of the message. In addition, Hill's (1973) findings suggest a significant tendency for subordinates to perceive their bosses as using one style of response to "handle interpersonal problems and another, different style to tackle technical problems" (p. 45).

In summary, results of investigations since 1970 that inquired into the nature of superior–subordinate feedback processes are generally consistent with conclusions from earlier research. Feedback from superiors to subordinates appears to be related to subordinate performance and satisfaction. However, at the same time, findings suggest that the subordinate's performance to a large extent controls the nature of his/her superior's feedback. Thus, present evidence indicates that future research should continue to explore the reciprocal character of feedback in superior–subordinate relationships, with particular emphasis on specific influence mechanisms.

Systemic Variables

Researchers have long been concerned with the effects that systemic organizational variables (e.g., technology, control structure, hierarchy, environment) have on the quality and nature of superior–subordinate communication. However, without doubt, less empirical as compared to theoretical research has been conducted in this area.

The exact role of technology in determining communication in organizations has puzzled researchers for at least two decades. For example, in the late 1950s Simpson (1959) found what he believed was the critical variable mediating the flow of vertical and horizontal communication in organizations: the degree of "mechanization" of work processes. Trist and Bamforth (1951), Gouldner (1954), Woodward (1958, 1965, 1970), Lawrence and Lorsch (1967), Perrow (1967), Pugh, Hickson, Hinings, and Turner (1969), and Peterson (1975), among others, all report findings which suggest that organizational members' perceptions of organizational and communication climate are linked to technological processes within organizations. Of even greater significance is Dubin's (1965) discovery that what is considered effective supervision may, in part, be a function of an organization's or work group's technology. Moreover, a recent doctoral study by Derry (1973) directly examined the effects technology and hierarchical position have on supervisory style, "communication responsiveness," and social interaction strategy. Results from two technologically diverse units within a manufacturing company (i.e., a manufacturing group and a research and development unit) indicated different patterns of superior–subordinate communication, depending on the interaction between technology and hierarchy.

Several investigations have also examined the relationship between organizational strucure and participant communication behavior (e.g., Bass, 1976; Blankenship & Miles, 1968; Ghiselli & Siegel, 1972; Porter & Lawler, 1964). Findings from these studies suggest the following conclusions that are relevant to superior–subordinate communication: (a) Upper level managers tend to involve their subordinates more in decision making than do lower level managers, whereas lower level managers tend to have decisions initiated for them by their superiors (Blankenship & Miles, 1968; Jago & Vroom, 1977); (b) "as compared with firms which have tall organizational structures, those which have flat structures reward with more rapid advancement those managers who favor sharing information and objectives with subordinates" (Ghiselli & Siegel, 1972, p. 622); and (c) in situations that are "regular, clear and structured," authoritative managerial direction is frequent, but subordinates often perceive such

direction "to be more effective under reverse conditions" (Bass, 1976, p. 215).

In summary, empirical research that has examined the relationship between systemic organizational variables and communication between superiors and subordinates has tended to focus on one of two areas: (a) technology or (b) organizational structure. However, the major portion of this research has often been simplistic and/or of limited generalizability. Future research in this area is sorely needed, for if we are to ever understand the micro system of superior–subordinate communication, we must first explicate its relationship with variables in the organization's macrosystem.

Discussion

This literature review has attempted to organize into nine topical categories empirical research on superior–subordinate communication. A close examination of this research suggests a number of conclusions. First, inspection of the variables explored within each category indicates that several basic constructs appear in more than one grouping. Specifically, we find at least three items that are consistently being studied: (a) the effects of power and status on superior–subordinate communication, (b) trust as a moderator of superior–subordinate communication, and (c) semantic-information distance as a source of misunderstanding in superior–subordinate communication. Moreover, these three variables tend to be explored concurrent to one another rather than in isolation of each other. Power and status differentials are an inevitable result of organizational development and in part serve as the impetus for the semantic-information distance between superiors and subordinates. However, the relationship of interpersonal trust to these variables is not perfectly clear, since in some situations it facilitates openness and understanding between superior and subordinates, whereas in other circumstances it appears to have no effect whatsoever. Moreover, there is some empirical and theoretical evidence (e.g., Sussman, 1975) which suggests that superior–subordinate semantic-information distance is a valuable and important feature of organization and that too large an attenuation of this gap may have dysfunctional consequences for organizational effectiveness. In short, the relationship between superior–subordinate semantic-information distance and organizational effectiveness may be a curvilinear association and may be one that is differentially moderated in various situations by interpersonal trust and perceived power and status differentials between the interactants. Obviously, additional multivariate research is required before we can fully understand the relationships between interpersonal trust, power and status, semantic-information distance, and superior–subordinate communication.

The second major conclusion to be drawn from the preceding review of literature is that a contingency/situational approach to the study of superior–subordinate communication is necessary. Repeatedly, we find that situational variables moderate the type and quality of communication exchanged between superiors and subordinates. Furthermore, if future research confirms the above proposition, much of the existing literature relating to superior–subordinate communication will become suspect and will require replication and clarification within and between organizations. In addition, it is essential that researchers start to explore the effects systemic organizational variables have on superior–subordinate communication, for as Graham and Roberts (1972) observe,

> we will only increase our understanding of organizational behavior as more researchers simultaneously investigate individual, group and organizational variables within organizations, and between organizations and the effects of environmental factors on those components. (p. 130)

Moreover, it is likely that such an approach will be more conducive to theory building in the area of superior–subordinate communication and organizational communication in general.

Finally, this review and interpretation of the literature identifies a need for some changes in the research questions we are asking about superior–subordinate communication and in the methods that are employed to answer those questions. At present, the majority of investigations exploring superior–subordinate communication have focused on

describing the various problematic states of superior–subordinate relations. For example, we can describe with a fair degree of confidence an open and closed superior–subordinate relationship, the communication qualities and characteristics of effective supervisors (for at least a limited number of situations), and the types of messages that tend to be distorted in upward communication. However, a much smaller amount of research has been directed towards discovering the antecedents to these conditions. For instance, we need to start asking questions such as How do open and closed superior-subordinate communication relationships develop? How do initial attributions and expectations of new superior–subordinate relations affect subsequent communication behavior? What stages of growth are characteristic of superior–subordinate communication relationships? How do superior–subordinate communication patterns change over time, and what causes these changes? And, of course, while these questions are being explored, a contingency orientation within our research investigations should be maintained. As Redding (1972) has observed, "the *contingency* nature of most generalizations about organizational phenomena should be kept in mind when interpreting the findings of studies dealing with supervisory communication" (p. 439).

The approach that Graen and his associates have employed to study dyadic linkages in superior–subordinate leadership exchanges provides an excellent model for the general study of superior–subordinate communication. Specifically, these researchers have attempted to trace the development of superior–subordinate relations from their initiation through the emergence of stable interaction patterns. In other words, they have used a developmental and longitudinal research design in exploring leadership behavior. Such an approach also appears to be ideal for the study of superior–subordinate communication, since it can provide descriptive, analytical, and often quasi-experimental data about superior–subordinate relationships. It should also be noted that a growing body of research is emerging that explores the relationships between applicant job expectations and employee attitudes, satisfaction, and so forth subsequent to employment in an organization (e.g., Ilgen & Seely, 1974; Katzell, 1968; Wanous, 1973). Information from these studies that relates to potential characteristics of superior–subordinate communication can serve as the initial point to begin our inquiries into the development of superior–subordinate communication patterns. In summary, combining our current knowledge of superior–subordinate communication with future studies that examine the dyad from a developmental, longitudinal perspective appears to provide the most promise for increased understanding of the phenomena we call superior–subordinate communication.

Reference Notes

1. Daly, J. A. *Communication and the personnel selection process: The state of the art*. Paper presented at the meeting of the International Communication Association, Chicago, April 1978.
2. Walton, E. *A magnetic theory of organizational communication* (Report No. 111). China Lake, Calif.: U.S. Naval Ordinance Test Station, January 1962.
3. Daly, J. A., McCroskey, J. C., & Falcione, R. L. *Communication apprehension, supervisor communication receptivity and satisfaction with superiors*. Paper presented at the meeting of the Eastern Communication Association, Philadelphia, Pa., April 1976.
4. Daly, J. A., McCroskey, J. C., & Falcione, R. L. *Homophily-heterophily and the prediction of supervisor satisfaction*. Paper presented at the meeting of the International Communication Association, Portland, Ore., April 1976.

References

Athanassiades, J. C. The distortion of upward communication in hierarchical organizations. *Academy of Management Journal*, 1973, *16*, 207–226.

Athanassiades, J. C. An investigation of some communication patterns of female subordinates in hierarchical organizations. *Human Relations*, 1974, *27*, 195–209.

Baird, J. W. An analytical field study of "open communication" as perceived by supervisors, subordinates, and peers (Doctoral dissertation, Purdue University, 1973). *Dissertation Abstracts International*, 1974, *35*, 562B. (University Microfilms No. 74–15, 116)

Baird, J. E., & Diebolt, J. C. Role congruence, communication, superior-subordinate relations and employee satisfaction in organizational hierarchies. *Western Speech Communication*, 1976, *40*, 260–267.

Barrow, J. C. Worker performance and task complexity as causal determinants of leader behavior. *Journal of Applied Psychology*, 1976, *61*, 433–440.

Bartol, K. M., & Butterfield, D. A. Sex effects in evaluating leaders. *Journal of Applied Psychology*, 1976, *61*, 446–454.

Bartol, K. M., & Wortman, M. S. Male versus female leaders: Effect on perceived leader behavior and satisfaction in a hospital. *Personnel Psychology*, 1975, *28*, 533–547.

Bass, B. M. A systems survey research feedback for management and organizational development. *Journal of Applied Behavioral Science*, 1976, *12*, 215–229.

Bass, B. M., Valenzi, E. R., Farrow, D. L., & Solomon, R. J. Management styles associated with organizational, task, personal, and interpersonal contingencies. *Journal of Applied Psychology*, 1975, *60*, 720–729.

Berkowitz, N. H., & Bennis, W. G. Interaction patterns in formal service-oriented organizations. *Administrative Science Quarterly*, 1961, *6*, 25–50.

Blankenship, V., & Miles, R. E. Organizational structure and managerial decision behavior. *Administrative Science Quarterly*, 1968, *13*, 106–120.

Bowers, D. G. Self-esteem and the diffusion of leadership style. *Journal of Applied Psychology*, 1963, *47*, 135–140.

Bowman, E. H. Consistency and optimality in managerial decision making. *Management Science*, 1963, *9*, 310–321.

Boyd, B. B., & Jensen, J. M. Perceptions of the first-line supervisor's authority: A study of superior-subordinate communication. *Academy of Management Journal*, 1972, *15*, 331–342.

Bradley, P. H., & Baird, J. E. Management and communicator style: A correlational analysis. *Central States Speech Journal*, 1977, *28*, 194–203.

Brenenstuhl, D. C. An empirical investigation of the leadership function as affected by leader style, interpersonal trust and commitment to future interaction (Doctoral dissertation, Indiana University, 1975). *Dissertation Abstracts International*, 1976, *36*, 6183A. (University Microfilms No. DAH 76-06, 027)

Brenner, M. H., & Sigband, N. B. Organizational communication—An analysis based on empirical data. *Academy of Management Journal*, 1973, *16*, 323–325.

Brown, D. S. Subordinate's views of ineffective executive behaviors. *Academy of Management Journal*, 1964, *7*, 288–299.

Browne, G. G., & Neitzel, B. J. Communication, supervision and morale. *Journal of Applied Psychology*, 1952, *36*, 86–91.

Burke, R. J. Methods of resolving superior-subordinate conflict: The constructive use of subordinate differences and disagreements. *Organizational Behavior and Human Performance*, 1970, *5*, 393–411.

Burke, R. J., & Wilcox, D. S. Effects of different patterns and degrees of openness in superior-subordinate communication on subordinate job satisfaction. *Academy of Management Journal*, 1969, *12*, 319–326.

Burke, R. J., Weir, T., & Duncan, G. Informal helping relationships in work organizations. *Academy of Management Journal*, 1976, *19*, 370–377.

Burnaska, R. F. The effects of behavior modeling training upon managers' behaviors and employees' perceptions. *Personnel Psychology*, 1976, *29*, 329–335.

Butler, R. P., & Jaffee, C. L. Effects of incentive, feedback, and manner of presenting the feedback on leader behavior. *Journal of Applied Psychology*, 1974, *59*, 332–336.

Cashman, J., Dansereau, F., Graen, G., & Haga, W. J. Organizational understructure and leadership: A longitudinal investigation of the managerial role-making process. *Organizational Behavior and Human Performance*, 1976, *15*, 278–296.

Clement, S. D. An analytical field study of selected message and feedback variables in the officer hierarchy of the U.S. Army (Doctoral dissertation, Purdue University, 1973) *Dissertation Abstracts International*, 1974, *35*, 609A. (University Microfilms No. 74-15, 144)

Close, J. M. Dogmatism and managerial achievement. *Journal of Applied Psychology*, 1975, *60*, 395–396.

Cohen, A. R. Upward communication in experimentally created hierarchies. *Human Relations*, 1958, *11*, 41–53.

Cook, D. M. The impact on managers of frequency of feedback. *Academy of Management Journal*, 1968, *11*, 263–278.

Dansereau, F., Cashman, J., & Graen, G. Instrumentality theory and equity theory as complementary approaches in predicting the relationship of leadership and turnover among managers. *Organizational Behavior and Human Performance*, 1973, *10*, 184–200.

Dansereau, F., Graen, G., & Haga, W. J. A vertical dyad linkage approach to leadership within formal organizations. *Organizational Behavior and Human Performance*, 1975, *13*, 46–78.

Day, D. R., & Stogdill, R. M. Leader behavior of male and female supervisors: A comparative study. *Personnel Psychology*, 1972, *25*, 353–360.

Dennis, H. S. *A theoretical and empirical study of managerial communication climate in complex organizations*. Unpublished doctoral dissertation, Purdue University, 1974.

Derry, J. D. A correlational and factor-analytic study of attitudes and communication networks in industry (Doctoral dissertation, Purdue University, 1972). *Dissertation Abstracts International*, 1973, *34*, 442A. (University Microfilms No. 73-15, 797)

Downs, C. W., & Pickett, T. An analysis of the effects of nine leadership-group compatability contingencies upon productivity and member satisfaction. *Communication Monographs*, 1977, *44*, 220–230.

Dubin, R. Supervision and productivity: Empirical findings and theoretical considerations. In R. Dubin (Ed.), *Leadership and productivity*. San Francisco: Chandler, 1965.

Dubin, R., & Spray, S. L. Executive behavior and interaction. *Industrial Relations*, 1964, *3*, 99–108.

Duffy, P. D. *Perceptions of satisfactory and unsatisfactory leadership styles of the junior high school principal*. Unpublished doctoral dissertation, University of Southern California, 1975.

Durand, D. E., & Nord, W. R. Perceived leader behavior as a function of personality characteristics of supervisors and subordinates. *Academy of Management Journal*, 1976, *19*, 427–438.

Eilon, E. Taxonomy of communication. *Administrative Science Quarterly*, 1968, *13*, 266–288.

Falcione, R. L. Credibility: Qualifier of subordinate participation. *Journal of Business Communication*, 1974, *11*, 43–54.

Fiedler, F. E. A contingency model of leadership effectiveness. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 1). New York: Academic Press, 1964.

Fiedler, F. E. *A theory of leadership effectiveness*. New York: McGraw-Hill, 1967.

Fiedler, F. E. Leadership experience and leadership performance: Another hypothesis shot to hell. *Organizational Behavior and Human Performance*, 1970, *5*, 1–14.

Fiedler, F. E. *Leadership*. New York: General Learning Press, 1971. (a)

Fiedler, F. E. Validation and extension of the contingency model of leadership effectiveness: A review of empirical findings. *Psychological Bulletin*, 1971, *76*, 128–148. (b)

Fiedler, F. E. The effects of leadership training and experience: A contingency model interpretation. *Administrative Science Quarterly*, 1972, *17*, 453–470. (a)

Fiedler, F. E. Predicting the effects of leadership training and experience from the contingency model. *Journal of Applied Psychology*, 1972, *56*, 114–119. (b)

Fleishman, E. A., & Harris, E. F. Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology*, 1962, *15*, 43–56.

Fleishman, E. A., Harris, E. F., & Burtt, H. E. *Leadership and supervision in industry*. Columbus, Ohio: Ohio State University, Bureau of Educational Research, 1955.

Fodor, E. M. Disparagement by a subordinate as an influence on the use of power. *Journal of Applied Psychology*, 1974, *59*, 652–655.

Funk, F. E. Communication attitudes of industrial foremen as related to their rated productivity (Doctoral dissertation, Purdue University, 1956). *Dissertation Abstracts*, 1956, *16*, 1015. (University Microfilms No. 00–16, 464)

Ghiselli, E. E., & Siegel, J. P. Leadership and managerial success in tall and flat organization structures. *Personnel Psychology*, 1972, *25*, 617–624.

Gibb, J. R. Defensive communication. *Journal of Communication*, 1961, *11*, 141–148.

Gouldner, A. W. *Patterns of industrial bureaucracy*. New York: Free Press, 1954.

Graen, G., Cashman, J. F., Ginsburgh, S. G., & Schiemann, W. Effects of linking-pin quality on the quality of working life of lower participants. *Administrative Science Quarterly*, 1977, *22*, 491–504.

Graham, W. K., & Roberts, K. H. (Eds.). *Comparative studies in organizational behavior*. New York: Holt, Rinehart, & Winston, 1972.

Green, S. G., Nebeker, D. M., & Boni, A. M. Personality and situational effects on leader behavior. *Academy of Management Journal*, 1976, *19*, 184–194.

Greene, C. N. Relationships among role accuracy, compliance, performance evaluation and satisfaction within managerial dyads. *Academy of Management Journal*, 1972, *15*, 205–216.

Greene, C. N. The reciprocal nature of influence between leader and subordinate. *Journal of Applied Psychology*, 1975, *60*, 187–193.

Greller, M. M., & Herold, D. M. Sources of feedback: A preliminary investigation. *Organizational Behavior and Human Performance*, 1975, *13*, 244–256.

Guetzkow, H. Communication in organizations. In J. G. March (Ed.), *Handbook of organizations*. Chicago: Rand McNally, 1965.

Haga, W. G., Graen, G., & Dansereau, F. Professionalism and role making in a service organization: A longitudinal investigation. *Administrative Science Quarterly*, 1974, *39*, 122–133.

Hall, J. Interpersonal style and the communication dilemma: I. Managerial implications of the Johari Awareness Model. *Human Relations*, 1974, *27*, 381–399.

Hall, J. Interpersonal style and the communication dilemma: II. Utility of the Johari Awareness Model for genotypic diagnosis. *Human Relations*, 1975, *28*, 715–736.

Haney, W. V. A comparative study of unilateral and bilateral communication. *Academy of Management Journal*, 1964, *7*, 128–136.

Haney, W. V. *Communication and organizational behavior—Text and cases* (2nd ed.). Homewood, Ill.: Irwin, 1967.

Harvey, J. B., & Boettger, C. R. Improving communication within a managerial workgroup. *Journal of Applied Behavioral Science*, 1971, *7*, 164–179.

Hawkins, B. L. Superior-subordinate communication as related to interpersonal need confirmation (Doctoral dissertation, Purdue University, 1975). *Dissertation Abstracts International*, 1976, *36*, 6366A. (University Microfilms No. DAH–76–07, 074)

Hegarty, H. W. Using subordinate ratings to elicit behavioral changes in supervisors. *Journal of Applied Psychology*, 1974, *59*, 764–766.

Heizer, J. H. Manager action. *Personnel Psychology*, 1972, *25*, 511–521.

Hill, W. A. Leadership style: Rigid or flexible. *Organizational Behavior and Human Performance*, 1973, *9*, 35–47.

Hill, W. A., & Hughes, D. Variations in leader behavior as a function of task type. *Organizational Behavior and Human Perfromance*, 1974, *11*, 83–96.

Hinrichs, J. R. Communications activity of industrial research personnel. *Personnel Psychology*, 1964, *17*, 193–204.

Hinton, B. L., & Barrow, J. C. The superior's reinforcing behavior as a function of reinforcements received. *Organizational Behavior and Human Performance*, 1975, *14*, 123–143.

House, R. L., Filley, A. C., & Gujarati, D. W. Leadership style, hierarchical influence, and the satisfaction of subordinate role expectations: A test of Likert's influence proposition. *Journal of Applied Psychology*, 1971, *55*, 422–432.

Housel, T. J., & Davis, W. E. The reduction of upward communication distortion. *Journal of Business Communication*, 1977, *14*, 49–65.

Ilgen, D. R., & Seely, W. Realistic expectations as an aid in reducing voluntary resignations. *Journal of Applied Psychology*, 1974, *59*, 452–455.

Indik, B. P., Georgopoulos, B. S., & Seashore, S. E. Superior-subordinate relationships and performance. *Personnel Psychology*, 1961, *14*, 357–374.

Jablin, F. M. An experimental study of message-response in superior-subordinate communication (Doctoral dissertation, Purdue University, 1977). *Dissertation Abstracts International*, 1978, *38*, 5796A. (University Microfilms No. 78–03, 241) (a)

Jablin, F. M. Message-response and "openness" in superior-subordinate communication. In B. D. Ruben (Ed.), *Communication yearbook II*. New Brunswick, N.J.: Transaction Books, 1978. (b)

Jago, A. G., & Vroom, V. H. Hierarchical level and leadership style. *Organizational Behavior and Human Performance*, 1977, *18*, 131–145.

Jain, H. C. Internal communications and supervisory effectiveness in two urban hospitals (Doctoral dissertation, University of Wisconsin—Madison, 1970). *Dissertation Abstracts International*, 1971, *31*, 5594A. (University Microfilms No. 71–03, 133)

Jain, H. C. Supervisory communication and performance in urban hospitals. *Journal of Communication*, 1973, *23*, 103–117.

Jones, A. P., James, L. R., & Bruni, J. R. Perceived leadership behavior and employee confidence in the leader as moderated by job involvement. *Journal of Applied Psychology*, 1975, *60*, 146–149.

Katz, D., & Kahn, R. *The social psychology of organizations*. New York: Wiley, 1966.

Katz, D., Maccoby, N., & Morse, N. C. *Productivity, supervision, and morale in an office situation*. Ann Arbor, Mich.: University of Michigan, Institute for Social Research, 1950.

Katzell, M. E. Expectations and dropouts in the schools of nursing. *Journal of Applied Psychology*, 1968, *52*, 154–158.

Kavanagh, M. J. Expected supervisory behavior, interpersonal trust and environmental preferences. *Organizational Behavior and Human Performance*, 1975, *13*, 17–30.

Kelly, C. M. "Actual listening behavior" of industrial supervisors, as related to "listening ability," general mental ability, selected personality factors and supervisory effectiveness (Doctoral dissertation, Purdue University, 1962). *Dissertation Abstracts*, 1963, *23*, 4019. (University Microfilms No. 63–02, 103)

Kelly, H. H. Communication in experimentally created hierarchies. *Human Relations*, 1951, *4*, 39–56.

Kelly, J. The study of executive behavior by activity sampling. *Human Relations*, 1964, *17*, 277–287.

Kerr, S., Schriesheim, C. A., Murphy, C. J., & Stogdill, R. M. Toward a contingency theory of leadership based upon the consideration and initiating structure literature. *Organizational Behavior and Human Performance*, 1974, *12*, 62–82.

Kim, J. S., & Hamner, C. W. Effects of performance feedback and goal setting on productivity and satisfaction in an organizational setting. *Journal of Applied Psychology*, 1976, *61*, 48–57.

Kipnis, D., & Lane, W. D. Self-confidence and leadership. *Journal of Applied Psychology*, 1962, *46*, 291–295.

Krivonos, P. D. Superior-subordinate communication as related to intrinsic and extrinsic motivation: An experimental field study (Doctoral dissertation, Purdue University, 1975). *Dissertation Abstracts International*, 1976, *36*, 4108A. (University Microfilms No. DAH 76–00, 552)

Lawler, E. E., Porter, L. W., & Tenenbaum, A. Managers' attitudes toward interaction episodes. *Journal of Applied Psychology*, 1968, *52*, 432–439.

Lawrence, P. R., & Lorsch, J. W. *Organization and environment: Managing differentiation and integration*. Boston, Mass.: Harvard University, Graduate School of Business Administration, Division of Research, 1967.

Leavitt, H. J., & Mueller, R. Some effects of feedback on communication. *Human Relations*, 1951, *4*, 401–410.

Level, D. A., & Johnson, L. Accuracy of information flows within the superior/subordinate relationship. *Journal of Business Communication*, 1978, *15*, 13–22.

Likert, R. *The human organization*. New York: McGraw-Hill, 1967.

Lowin, A., & Craig, J. R. The influence of level of performance on managerial style: An experimental object-lesson in the ambiguity of correlational data. *Organizational Behavior and Human Performance*, 1968, *3*, 440–458.

MacDonald, D. Communication roles and communication networks in a formal organization. *Human Communication Research*, 1976, *2*, 365–375.

Maier, N. R. F., Hoffman, R. L., & Read, W. H. Superior-subordinate communication: The relative effectiveness of managers who held their subordinate's positions. *Personnel Psychology*, 1963, *16*, 1–11.

Maier, N. R. F., Hoffman, R. L., Hooven, J. L., & Read, W. H. Superior-subordinate communication: A statistical research project. *American Management Research Studies*, 1961, *52*, 9–30.

Marcus, P. M., & House, J. S. Exchange between superiors and subordinates in large organizations. *Administrative Science Quarterly*, 1973, *18*, 209–222.

Melcher, A. J., & Beller, R. Toward a theory of organizational communication: Consideration in channel selection. *Academy of Management Journal*, 1967, *10*, 39–52.

Mellinger, G. D. Interpersonal trust as a factor in communication. *Journal of Abnormal Social Psychology*, 1956, *52*, 304–309.

Meyer, H. H., Kay, E., & French, J. R. P. Split roles in performance appraisal. *Harvard Business Review*, 1965, *43*, 123–129.

Miles, R. E. Attitudes toward management theory as a factor in managers' relationships with their superiors. *Academy of Management Journal*, 1964, *7*, 308–314.

Minter, R. L. A comparative analysis of managerial communication in two divisions of a large manufacturing company (Doctoral dissertation, Purdue University, 1969). *Dissertation Abstracts International*, 1969, *30*, 2653A. (University Microfilms No. 69–17, 221)

Miraglia, J. F. An experimental study of the effects of communication training upon perceived job performance of nursing supervisors in two urban hospitals (Doctoral dissertation, Purdue University, 1963). *Dissertation Abstracts*, 1964, *24*, 5611. (University Microfilms No. 64–05, 749)

Mitchell, T. R., Smyser, C. M., & Weed, S. E. Locus of control: Supervision and work satisfaction. *Academy of Management Journal*, 1975, *18*, 623–631.

Moore, M. L. Superior, self, and subordinate differences in perceptions of managerial learning times. *Personnel Psychology*, 1974, *27*, 297–305.

Northouse, P. G. Predictors of empathic ability in an organizational setting: A research note. *Human Communication Research*, 1977, *3*, 176–178.

Odiorne, G. S. An application of the communication audit. *Personnel Psychology*, 1954, *7*, 235–243.

Oldham, G. R. The motivational strategies used by superiors: Relationships to effectiveness indicators. *Organizational Behavior and Human Performance*, 1976, *15*, 66–86.

O'Reilly, C. A., & Roberts, K. H. Information filtration in organizations: Three experiments. *Organizational Behavior and Human Performance*, 1974, *11*, 253–265.

Organ, D. W. Social exchange and psychological reactance in a simulated superior-subordinate relationship. *Organizational Behavior and Human Performance*, 1974, *12*, 132–142.

Pelz, D. C. Influence: A key to effective leadership in the first-line supervisor. *Personnel*, 1952, *29*, 3–11.

Penfield, R. V. Time allocation patterns and effectiveness of managers. *Personnel Psychology*, 1974, *27*, 245–255.

Perrow, C. A framework for the comparative analysis of organizations. *Administrative Science Quarterly*, 1967, *32*, 194–208.

Peterson, R. B. The interaction of technological process and perceived organizational climate in Norwegian firms. *Academy of Management Journal*, 1975, *18*, 288–299.

Petty, M. M., & Lee, G. L. Moderating effects of sex of supervisor and subordinate on relationships between supervisory behavior and subordinate satisfaction. *Journal of Applied Psychology*, 1975, *60*, 624–628.

Petty, M. M., & Miles, R. H. Leader sex-role stereotyping in a female dominated work culture. *Personnel Psychology*, 1976, *29*, 393–404.

Pfeffer, J., & Salancik, G. R. Determinants of supervisory behavior: A role set analysis. *Human Relations*, 1975, *28*, 139–154.

Pinder, C. C., & Pinto, P. R. Demographic correlates of managerial style. *Personnel Psychology*, 1974, *27*, 257–270.

Ponder, Q. D. Supervisory practices of effective and ineffective foremen (Doctoral dissertation, Columbia University, 1958). *Dissertation Abstracts*, 1959, *20*, 3983. (University Microfilms No. 59–01, 497)

Porter, L. W. Differential self-perceptions of management personnel and line workers. *Journal of Applied Psychology*, 1958, *42*, 105–108.

Porter, L. W., & Lawler, E. E. The effects of "tall" versus "flat" organization structures on managerial job satisfaction. *Personnel Psychology*, 1964, *17*, 135–148.

Pugh, D. S., Hickson, D. J., Hinings, C. R., & Turner, C. The content of organization structures. *Administrative Science Quarterly*, 1969, *14*, 91–114.

Pyron, H. C. The construction and validation of a forced-choice scale for measuring oral communication attitudes of industrial foremen (Doctoral dissertation, Purdue University, 1964). *Dissertation Abstracts*, 1964, *25*, 1413. (University Microfilms No. 64–08, 702)

Read, W. H. Upward communication in industrial hierarchies. *Human Relations*, 1962, *15*, 3–15.

Redding, W. C. *Communication within the organization: An interpretive review of theory and research*. New York: Industrial Communication Council, 1972.

Renwick, P. A. Perception and management of superior-subordinate conflict. *Organizational Behavior and Human Performance*, 1975, *13*, 444–456.

Richetto, G. M. Source credibility and personal influence in three contexts: A study of dyadic communication in a complex aerospace organization (Doctoral dissertation, Purdue University, 1969). *Dissertation Abstracts International*, 1969, *30*, 1668A. (University Microfilms No. 69–17, 245)

Rizzo, J. R., House, R. J., & Lirtzmann, S. I. Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly*, 1970, *15*, 150–163.

Roberts, K. H., & O'Reilly, C. A. Failures in upward communication: Three possible culprits. *Academy of Management Journal*, 1974, *17*, 205–215.

Rogers, E. M., & Rogers, R. A. *Communication in organizations*. New York: Free Press, 1976.

Rosen, B., & Adams, J. S. Organizational coverups: Factors influencing the discipline of information gatekeepers. *Journal of Applied Social Psychology*, 1974, *4*, 375–384.

Rosen, B., & Jerdee, T. H. The influence of sex role stereotypes on evaluations of male and female supervisory behavior. *Journal of Applied Psychology*, 1973, *57*, 44–48.

Rosen, H. Managerial role interaction: A study of three managerial levels. *Journal of Applied Psychology*, 1961, *45*, 30–34.

Rosen, S., & Tesser, A. On reluctance to communicate undesirable information: The MUM effect. *Sociometry*, 1970, *33*, 253–263.

Rubin, I. M., & Goldman, M. An open system model of leadership performance. *Organizational Behavior and Human Performance*, 1968, *3*, 143–156.

Sadler, D. J. Leadership style, confidence in management and job satisfaction. *Journal of Applied Behavioral Science*, 1970, *6*, 3–19.

Sank, L. I. Effective and ineffective managerial traits obtained as naturalistic descriptions from executive members of a super corporation. *Personnel Psychology*, 1974, *27*, 423–434.

Schein, V. E. The relationship between sex role stereotypes and requisite management characteristics. *Journal of Applied Psychology*, 1973, *57*, 95–100.

Schein, V. E. Relationships between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology*, 1975, *60*, 340–344.

Schriesheim, C. A., & Murphy, C. J. Relationships between leader behavior and subordinate satisfaction and performance: A test of some situational moderators. *Journal of Applied Psychology*, 1976, *61*, 634–641.

Schwartz, M. M., Stark, H. F., & Schiffman, H. R. Responses of union and management leaders to emotionally toned industrial relations terms. *Personnel Psychology*, 1970, *23*, 361–367.

Senger, J. Managers' perceptions of subordinates' competence as a function of personal value orientations. *Academy of Management Journal*, 1971, *14*, 415–423.

Simons, H. W. A comparison of communication attributes and rated job performance of supervisors in a large commercial enterprise (Doctoral dissertation, Purdue University, 1961). *Dissertation Abstracts*, 1962, *22*, 3778. (University Microfilms No. 62-00, 887)

Simpson, R. L. Vertical and horizontal communication in formal organizations. *Administrative Science Quarterly*, 1959, *4*, 188–196.

Sims, H. P., & Szilagyi, A. D. Leader reward behavior and subordinate satisfaction and performance. *Organizational Behavior and Human Performance*, 1975, *14*, 426–438.

Sincoff, M. Z. An experimental study of the effects of three "interview styles" upon judgments of the interviewees and observer-judges (Doctoral dissertation, Purdue University, 1969). *Dissertation Abstracts International*, 1970, *30*, 5100A. (University Microfilms No. 70–08, 972)

Smith, C. G., & Kight, S. S. Effects of feedback on insight and problem solving efficiency in training groups. *Journal of Applied Psychology*, 1959, *43*, 209–211.

Smith, R. L. Communication correlates of interpersonal sensitivity among industrial supervisors (Doctoral dissertation, Purdue University, 1967). *Dissertation Abstracts*, 1968, *28*, 4743A. (University Microfilms No. 68–06, 361)

Stogdill, R. M. *Handbook of leadership*. New York: Free Press, 1974.

Stull, J. B. "Openness" in superior-subordinate communication: A quasi-experimental field study (Doctoral dissertation, Purdue University, 1974). *Dissertation Abstracts International*, 1975, *36*, 603A. (University Microfilms No. 75–17, 285)

Sussman, L. Upward communication in the organizational hierarchy: An experimental field study of perceived message distortion (Doctoral dissertation, Purdue University, 1973). *Dissertation Abstracts International*, 1974, *34*, 5366A. (University Microfilms No. 74–05, 055)

Sussman, L. Communication in organizational hierarchies: The fallacy of perceptual congruence. *Western Speech Communication*, 1975, *39*, 191–199.

Sussman, L., Pickett, T. A., Berzinski, I. A., & Pearce, F. W. Sex and sycophancy: Communication strategies for ascendance in same sex and mixed sex superior-subordinate dyads. *Sex Roles*, in press.

Tacey, W. S. *Critical requirements for the oral communication of industrial foremen*. Pittsburgh, Pa.: Author, 1959.

Tenenbaum, A. Dyadic communications in industry (Doctoral dissertation, University of California, Berkeley, 1970). *Dissertation Abstracts International*, 1971, *31*, 7662B. (University Microfilms No. 71–15, 902)

Tompkins, P. K. *An analysis of communication between headquarters and selected units of a national labor union*. Unpublished doctoral dissertation, Purdue University, 1962.

Tosi, H. L. The effect of the interaction of leader behavior and subordinate authoritarianism. *Personnel Psychology*, 1973, *26*, 339–350.

Triandis, H. C. Categories of thought of managers, clerks and workers about jobs and people in an industry. *Journal of Applied Psychology*, 1959, *43*, 338–344. (a)

Triandis, H. C. Cognitive similarity and interpersonal communication in industry. *Journal of Applied Psychology,* 1959, *43,* 321–326. (b)

Triandis, H. C. Differential perceptions of certain jobs and people by managers, clerks and workers in industry. *Journal of Applied Psychology,* 1959, *43,* 221–225. (c)

Triandis, H. C. Cognitive similarity and communication in a dyad. *Human Relations,* 1960, *13,* 175–183.

Trist, E. L., & Bamforth, K. W. Some social and psychological consequences of the longwall method of coal-getting. *Human Relations,* 1951, *4,* 3–38.

Vroom, V. H. Industrial social psychology, In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 5). Reading, Mass.: Addison-Wesley, 1970.

Vroom, V. H. Leadership. In M. C. Dunnette (Ed.), *Handbook of industrial and organizatinal psychology.* Chicago: Rand McNally, 1976.

Vroom, V. H., & Mann, F. C. Leader authoritarianism and employee attitudes. *Personnel Psychology,* 1960, *13,* 125–140.

Wager, L. W. Leadership style, influence, and supervisory role obligations. *Administrative Science Quarterly,* 1965, *9,* 391–420.

Walker, C. R., Turner, A. N., & Guest, R. H. *The foreman on the assembly line.* Cambridge, Mass.: Harvard University Press, 1956.

Walter, B. Internal control relations in administrative hierarchies. *Administrative Science Quarterly,* 1966, *11,* 179–206.

Wanous, J. P. Effects of a realistic job preview on job acceptance, job attitudes, and job survival. *Journal of Applied Psychology,* 1973, *58,* 327–332.

Weaver, C. H. The quantification of the frame of reference in labor-management communication. *Journal of Applied Psychology,* 1958, *42,* 1–9.

Webber, R. A. Perceptions of interactions between superiors and subordinates. *Human Relations,* 1970, *23,* 235–248.

Weed, S. E., Mitchell, T. R., & Moffitt, W. Leadership style, subordinate personality, and task type as predictors of performance and satisfaction with supervision. *Journal of Applied Psychology,* 1976, *61,* 58–66.

White, B. Y. Superordinate and subordinate perceptions of managerial styles of selected male and female college administrators (Doctoral dissertation, Brigham Young University, 1976). *Dissertation Abstracts International,* 1976, *37,* 841A. (University Microfilms No. DAH76–18, 355)

White, H. C. Perceptions of leadership by managers in a federal agency. *Personnel Administration/Public Personnel Review,* 1972, *1,* 51–56.

Willits. R. D. Company performance and interpersonal relations. *Industrial Management Review,* 1967, *7,* 91–107.

Woodward, J. *Management and technology.* London. Her Majesty's Stationery Office, 1958.

Woodward, J. *Industrial organization: Theory and practice.* London: Oxford University Press, 1965.

Woodward, J. *Industrial organization: Behaviour and control.* London: Oxford University Press, 1970.

Yoder, D. *Personnel management and industrial relations* (6th ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Young, J. W. The subordinate's exposure of organizational vulnerability to the superior: Sex and organizational effects. *Academy of Management Journal,* 1978, *21,* 113–122.

Zajonc, R. B. The effects of feedback and probability of group success on individual and group performance. *Human Relations,* 1962, *15,* 149–161.

Zima, J. P. The counseling-communication of supervisors in a large manufacturing company (Doctoral dissertation, Purdue University, 1968). *Dissertation Abstracts,* 1969, *29,* 3956B. (University Microfilms No. 69–07, 518)

Received June 1, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1223–1235

Clinical Applications of Hypnosis to Three Psychosomatic Disorders

Frank A. De Piano and Herman C. Salzberg
University of South Carolina

Studies of hypnosis in the treatment of skin disorders, headaches, and asthma were reviewed in terms of outcomes and methodological soundness. Some studies focused on changing physiological functions, others on increasing insight in their patients, and still others on altering patients' perceptions of their symptoms. Methodological weaknesses included lack of control groups, nonrandom assignment of patients to treatment conditions, and confounding of treatment effects or lack of control for placebo effects. Additional weaknesses centered around the use of single outcome measures and the failure to assess the specific roles of mediating variables. Most of the studies reviewed showed positive treatment effects. However, there is equivocal evidence that hypnosis can directly influence autonomic functioning. Hypnosis may be valuable in facilitating one's capacity to gain insight into how one's symptoms developed and are maintained. In addition, hypnotic procedures have resulted in some success when used to indirectly alleviate symptoms by altering how individuals perceive their disorders and how these disorders affect their lives.

Hypnosis has become an accepted mode of treatment in both medicine and clinical psychology. Its judicious use was endorsed by the American Medical Association in 1958 and by the British Medical Association in 1955. In 1960, the American Psychological Association sanctioned its use in clinical practice (Hilgard, 1968). Several societies are devoted to the professional study of hypnosis, and both psychology and medical training programs offer courses in it.

This article reviews the published reports of the clinical application of hypnosis to skin disorders, headaches, and asthma. The review covers the period from 1967 to 1977 comprehensively, as well as reports prior to 1967 that had a substantial impact on the field.

Hypnosis is most frequently associated with suggestibility. Barber (1974) and his collaborators have, in essence, equated the two by arguing that subjects, when given direct suggestions without a hypnotic induction, displayed behavior similar to that of subjects given a formal induction. Other investigators have adhered to the conceptualization of hypnosis as an altered state of consciousness (Fromm, 1977; Hilgard, 1968). In either case, its investigation as a treatment adjunct for psychosomatic disorders would be justifiable. However, expectations for success might vary with the definition that the investigator accepts.

Although there is disagreement about which specific ailments are psychosomatic and which are primarily organic, it is widely accepted that psychosomatic disorders are etiologically related to or are exacerbated by psychological factors, even though physical symptoms are present and medical intervention is often necessary. Some practitioners argue that illness is always a psychosomatic event. Others merely label a disorder psychosomatic when physical causes are not readily identifiable. Obviously, an investigator's view of a disorder as psychosomatic would influence the decision to use hypnosis in a particular case, which might, in turn, affect the outcome of treatment.

Requests for reprints should be sent to Herman C. Salzberg, Department of Psychology, University of South Carolina, Columbia, South Carolina 29208.

Most investigators do not present theoretical arguments for using hypnosis with psychosomatic disorders. Therefore, much must be inferred. From the theories that are discussed in the literature, three rationales seem to emerge. First, there are those who contend that through hypnosis and hypnotic suggestions, certain autonomic nervous system functions, which are not readily controlled by the individual, come under voluntary control and mediate symptom changes. For example, Clawson and Swade (1975) contend that blood flow to internal organs and limbs can be controlled.

Next, there are investigators who use hypnosis for its assumed capacity to assist patients in gaining insight into how their symptoms develop and are maintained. Once this insight is achieved, it is further assumed that individuals will be able to solve their problems, thus eliminating the symptoms.

Finally, there are investigators who use hypnosis to alter the way in which one's disorder is perceived. For example, suggestions are given that the symptoms will no longer concern the patient. This altered perception has two potential effects. First, the patient can become less debilitated by these symptoms and thereby can function more effectively, and second, the emotional component of the disorder, which may exacerbate the condition, can be reduced.

Research articles have occasionally appeared that investigate the use of hypnosis in treatment of other psychosomatic disorders, such as Raynaud's disease, arthritis, seizures, stomach disorders, Gilles de la Tourette's syndrome, hypertension, and so forth. Hypnosis research has been reported most frequently in the treatment of skin disorders, headaches, and asthma. Therefore, the present review has been limited to these three disorders.

Methodological Considerations

As is true of much clinical research, studies of hypnosis in the treatment of psychosomatic disorders reflect many methodological shortcomings. Understandably, in a clinical setting, it is frequently extremely difficult to establish proper controls. Considerably more than half of the studies reviewed were single- or multiple-case reports. A few used a treatment group procedure but did not include a comparison group. Of the remaining controlled studies, some did not randomly assign subjects to treatment conditions.

A glaring weakness across the majority of studies was the lack of specification of the hypnotic procedure used. Few studies discussed the induction technique in detail, and even fewer reported the use of a reliable instrument for assessing whether the patient was hypnotized. An additional widespread shortcoming was the lack of control for experimenter bias, which could have been accomplished by keeping the experimenter blind to patient treatment group.

Methodological weaknesses were also noted in terms of the outcome measures employed. Self-report and physicians' impressions were most frequently used. Physiological change was assessed on occasion. A combination of both objective and subjective measures, however, was infrequently reported.

Finally, none of the studies that were reviewed employed an acceptable control for placebo effects.

A summary of the studies can be found in Table 1, which includes an indication of whether the treatment was successful (+) or unsuccessful (−), time interval of follow-up, type of design (case, treatment group, treatment and control condition), whether or not subjects were randomly assigned, and the number of patients/subjects studied.

Skin Disorders

Hypnosis is commonly used as one form of treatment for skin disorders because of the generally assumed association of skin responses and the autonomic nervous system. Both the skin and the autonomic nervous system develop embryologically from the ectoderm, and correlations between self-report of the normal responses to embarrassment with blushing and to fear with blanching and goose pimples are well accepted (Jabush, 1969). The types of disorders treated include psoriasis (a condition characterized by reddish and silvery scales), dermatitis (an inflamation of an area of the skin), boils (pyo-

Table 1
Experimental Designs and Treatment Outcomes

Study	Results[a] P	I	S	Follow-up	Design	Random Assignment	*n*
				Skin Disorders			
Physiological change							
Twerski & Naar (1974)		+	+	6 months	Case report	na	1
Ewin (1974)		+	+	6 years	Case report	na	1
Jabush (1969)		+	+	2½ years	Case report	na	1
Tasini & Hackett (1977)		+	+	4–8 months	Case report	na	3
Clawson & Swade (1975)		+	+	0–4 years	Case report	na	3
Vollmer (1946)		+	+	Not reported	Multiple case	na	7
Surman, Gottlieb, Hackett, & Silverberg (1973)		+	+	3 months	Treatment group	No	24
Peters & Stern (1971)	−			na	Treatment and control conditions	Yes	20
Johnson & Barber (1976)	−	−		na	Treatment and control conditions	Yes	48
Insight							
Ewin (1974)		+	+	1½–5 years	Multiple case	na	3
French (1973)		+	+	5 weeks	Case report	na	1
Frankel & Misch (1973)		+	+	13 months	Case report	na	1
Altered perception							
Ament & Milgrom (1967)		+	+	2 months	Case report	na	1
Klinge (1971)		+	+	Not reported	Multiple case	na	3
				Headache			
Physiological change							
Ansel (1977)		+	+	Not reported	Case report	na	1
Todd & Kelly (1970)		+	+	1 month	Case report	na	1
Graham (1975)		+	+	9–12 months	Case report	na	2
Harding (1961)		+	+	6–30 months	Treatment group	No	25
Cedercreutz, Lahteenmaki, & Tulikoura (1976)		+	+	22 months	Treatment group	No	155
Andreychuk & Skriver (1975)		−	−	None	Treatment and control condition	Yes	33
Anderson, Basker, & Dalton (1975)		+	+	1 year	Treatment and control condition	Yes	47
Insight							
Blumenthal (1963)		+	+	1 year	Case report	na	2
Altered perception							
Kroger (1963)		+	+	Not reported	Multiple case	na	Not reported
				Asthma			
Physiological change							
Dennis (1965)		+	+	na	Treatment group	No	5
Weiss, Martin, & Riley (1970)		−		na	Treatment and control condition	Yes	16
Smith & Burns (1960)	−	−		4 weeks	Treatment group	No	25
Mun (1969)	+	+	+	Not reported	Treatment group	No	10
Aronoff, Aronoff, & Peck (1975)	+	+		Not reported	Treatment group	No	17
Philipp, Wilde, & Day (1972)	+	+	+	na	Treatment group	No	20
Fry, Mason, & Pearson (1964)	+			na	Treatment and control condition	Yes	47
Altered perception							
Hanley (1974)	+			6–30 months	Case report	na	2
Smith (1970)	+	+	+	na	Case report	na	2
Edwards (1960)	−	+	+	1 year	Treatment group	na	6
Mun (1969)	+	+		6 months	Treatment group	No	36
Moorefield (1971)	+	+		17 months	Treatment group	No	9
Collison (1975)	+	+		Not reported	Treatment group	No	121
White (1961)	−	+		12–17 months	Treatment group	No	10
Maher-Laughan (1970)							
Study A		+		6 months	Treatment and control condition	Yes	55
Study B		+		1 year		Yes	252
Study C		+		6 years	Treatment group	No	173

Note. P = physiological or objective change associated with symptom removal; I = physician's or investigator's impression of symptom removal; S = self-report of symptom remission; na = not applicable.
[a] + indicates that treatment was successful; − indicates that treatment was unsuccessful.

genic infections originating in hair follicles), and warts (thickenings of the Malpighian and granular layers of the epidermis).

Physiological changes. A number of single- and multiple-case studies have reported the successful application of hypnosis to various skin disorders, for example, Twerski and Naar (1974), refractory dermatitis; Jabush (1969), boils; and Tasini and Hackett (1977), Clawson and Swade (1975), Vollmer (1946), and Ewin (1974), warts. These investigations did not take into account baseline rates of spontaneous remission, and in addition, they often confounded the effects of hypnosis with other treatment effects. Although several investigators hypothesized that autonomic changes would occur through hypnosis, no physiological measurements were made.

Surman, Gottlieb, Hackett, and Silverberg (1973) conducted a controlled investigation of the effects of hypnotic treatment of warts. Seventeen patients were told, during each weekly hypnotic session, that they would experience a tingling sensation in the warts on one side of their body and that only those warts would subsequently disappear. Treatment continued for 5 consecutive weeks. Seven patients served as a waiting list control group. The authors concluded that warts respond to hypnosis but that hypnotic suggestion does not affect them selectively. There were, however, critical differences between the treatment and comparison groups. The average duration since the appearance of the warts had been 2.8 and 3.9 years, respectively, for these groups. As reported by the authors, warts spontaneously disappeared, on the average, 2.28 years after onset. It may well be that the warts of the patients in the comparison group were resistant to remission over time, whereas those in the treatment group were less resistant. The conclusion that hypnotic suggestion can eradicate warts becomes even more suspect when, although none of the 7 control subjects lost any warts during the 3-month treatment period, one of them, who was scheduled for postexperiment hypnotherapy, spontaneously lost all but one wart before the first treatment session.

The few well-controlled studies that investigated physiological changes were not performed on clinical populations. In a laboratory study conducted by Peters and Stern (1971), it was hypothesized that blood pressure, skin temperature, and pulse rate would increase in subjects given suggestions of hives and would decrease in those given the suggestion that they would show symptoms (peripheral vasoconstriction) associated with Raynaud's disease. These effects had been observed during the natural occurrence of these disorders. Subjects were well screened for hypnotic susceptibility before entry into the experiment. There were four treatment groups, each of which received two sessions. Group 1 received suggestions for hives while hypnotized in Session 1 and suggestions for symptoms of Raynaud's disease with method acting instructions in Session 2. Group 2 received the same treatment as Group 1, with sessions given in reverse order. Group 3 received suggestions for symptoms of Raynaud's disease while hypnotized, then hives suggestions with method acting instructions. Group 4 was treated identically to Group 3, with sessions administered in reverse order. Although Peters and Stern did not elaborate on what they meant by method acting, it appears to be similar to role playing. Results did not support the contention that specific changes would occur depending on suggestions given during hypnosis. It was observed, however, that hypnotized subjects, regardless of the suggestions given, did show a decrease in skin temperature and blood volume in the fingers.

In a second well-controlled nonclinical experiment, conducted by Johnson and Barber (1976), the effects of hypnotic suggestion on blister formation were examined. This study was especially noteworthy for its attempt to control many of the factors that possibly influenced results of earlier studies, including (a) direct observation of subjects, to eliminate the possibility that skin changes might be a function of self-injurious behavior; (b) procedures to insure that the experimenter remained blind, to reduce the possibility of experimenter bias; (c) an evaluation of hypnotic susceptibility; and (d) an evaluation of skin sensitivity, to assess its influence on skin change. An additional strong point of this study was the assessment of skin temperature

changes, which were assumed to mediate blister formation, in addition to the direct observation of external changes of the skin.

None of the experimental subjects displayed any true blisters. In addition, only 2 of the 48 subjects displayed visible skin changes. In both of these subjects, changes could be attributed to factors other than hypnotic suggestion. No differences in temperature were found between the test hand and control hand. Comparison of susceptible versus nonsusceptible subjects revealed no differential change in hand temperature. The actual data were not presented, however, making it impossible to determine if this nonsignificant finding was due to a lack of a cause–effect relationship or because of a lack of statistical power, due to the small number of subjects who were considered susceptible. Although the study indicated that the general population may not show physical changes in response to suggestion, it does not preclude the possibility that subjects who spontaneously develop blisters (and other skin disorders) might respond positively to suggestion.

Insight. Several case studies illustrate how hypnosis was used to help patients gain insight into the development and maintenance of their symptoms.

Ewin (1974) reported on three cases of successful treatment of conyloma acumatum (warts on the genitalia or perineal region). The first case met with initial failure when suggestions for direct physiological changes were made. After this failure, the assumption was made that the warts were serving some purpose not consciously perceived by the patient. While hypnotized, the patient revealed that he had been engaging in extramarital intercourse, which he thought might be having a deleterious effect on his marriage. The warts were restricting this extramarital activity, thereby bringing the patient closer to his wife. During hypnosis, it was suggested that this type of problem solving was inefficient. Within 3 weeks of this single session, all the warts disappeared. In a second similar case, a male patient revealed, while hypnotized, that his warts kept him from engaging in extramarital relationships. He was told that his behavior could be consciously controlled and that the warts were unnecessary. Two-and-one-half months later, the warts were two thirds gone, with complete remission occurring 1 month later. In a third case, suggestions without hypnosis, were given to a male homosexual with a rosette of perineal warts. It was suggested to the patient that he feared anal rape, and that he should forget about protecting his anus with this ring of warts. One month later, the author reported 50% remission, and after one additional session with hypnosis, complete remission was claimed. In each of these cases, length of time between initial treatment and eventual cure ranged from 2 to 3½ months. Spontaneous remission of symptoms might have occurred during that time.

French (1973) described the successful treatment of a woman who suffered from venereal warts. While hypnotized, the patient recalled that the warts first appeared during a gratifying sexual affair. During this session the patient was told that she had to choose between keeping both the affair and the warts or relinquishing the affair and thereby the warts. The woman chose to give up the affair and the warts. After 3 weeks it was reported that she was greatly improved, with complete recovery noted after 5 weeks.

Frankel and Misch (1973) reported on a case of long-standing psoriasis in which hypnosis was used to help a patient gain insight. After several sessions that used hypnotic suggestion to influence symptoms directly, the patient was requested to discuss his resistance to relinquishing his psoriasis. He stated that he realized that maintenance of the symptom allowed him the comfort of avoiding threatening social interactions. After 8 weeks, the patient had improved considerably. The authors indicated that in addition to the self-report and observations of improvement, there was also a 2.1 °F. (1.17 °C) increase in skin temperature from the beginning of the trance to its completion. It was impossible to determine, however, whether this change was a function of hypnosis, suggestion alone, emotional changes, or a byproduct of a spontaneous remission of the condition.

Altered perception. This final pair of investigations employed hypnosis with skin disorders to alter the subjective perception and/or reduce the discomfort created by the symptoms.

Ament and Milgrom (1967) found hypnotic suggestion to be successful in the treatment of a case of pruritus. After the first session, decreased pruritus was reported with additional improvement noted after each subsequent session. After five sessions, however, a previously suspected diagnosis of myelocytic leukemia was made. For the following 2 months, the patient's skin remained generally in good condition. It cannot be known with certainty, however, what psychological effect the development of leukemia may have had on the patient and consequently on the pruritus condition.

In a second series of cases presented by Klinge (1971), three dermatitis patients were successfully treated with hypnotic suggestions and subsequently perceived their symptoms as less severe. Unfortunately, this treatment effect was confounded by simultaneous treatment with vitamin A, hydrocortisone, and bandaging.

It appears that well-controlled studies of skin disorders have not substantiated the case study findings of physical change associated with the hypnotic treatment. Evidence for autonomic changes, which are often assumed to mediate external skin changes, was typically absent, or the autonomic changes were not measured. Although inconclusive because of the lack of control for spontaneous remission, there were several interesting instances of symptom improvement as a result of the use of hypnosis to help patients gain insight into the maintenance of their symptoms. Some success, although limited to a small number of cases, was also reported as a result of attempts to alter perceptions and reduce the discomfort created by these symptoms.

Headaches

A second disorder that is commonly treated with hypnosis and hypnotic suggestion is headache. Migraine and tension headaches are most commonly treated. Migraine headaches are thought to be initiated by constriction of the blood vessels to the head and brain, which is followed by overcompensation, causing the vessels to become enlarged. This expansion is credited with causing the unilateral pain, dizziness, nausea, and blurred vision associated with migraine (Wolff, 1963). Traditional treatment consists of two interventions, one aimed at inhibiting the initial constriction of the vessels, the other aimed at reducing the dilation of blood vessels after onset of the headache (Graham, 1975). Tension headaches, on the other hand, are a result of prolonged muscular contractions (usually associated with stress) about the face, scalp, and neck. Behavioral treatment generally consists of deep muscle relaxation training or biofeedback.

Physiological changes. Successful treatment of individual cases of migraine headache has been reported by Ansel (1977) and Graham (1975) and of tension headaches by Todd and Kelly (1970). Treatment of 25 cases of intractable migraine was reported by Harding (1961). All of the patients' symptoms had become progressively worse despite previous drug treatments. Hypnosis was recommended as a last resort. Treatment consisted of four sessions that included history taking, explanation of migraine, and hypnotic suggestions that the patients picture the blood vessels in their heads growing smaller and returning to normal. Improvement was assessed in two ways: through self-report and through the hypnotist's impressions. Five had complete remission, 10 showed substantial improvement, and 5 experienced some relief. Five patients showed no improvement at all. Three of the 5 failures were reported not to be hypnotizable. This study is noteworthy for the effort made to systematically treat a number of patients with a standard approach. However, no report of the successful patients' hypnotizability was made, thus preventing an assessment of the relationship between hypnotizability and treatment success.

Cedercreutz, Lahteenmaki, and Tulikoura (1976) treated 155 consecutive skull-injured patients who had been experiencing headache and vertigo for 1 week or longer after their injury. Treatment consisted of between 1 and 10 weekly sessions. During hypnosis, therapeutic suggestions were repeated several times.

Trance capability was assessed by the investigators on a 4-point rating scale. The authors reported that although 58% of the 120 headache patients who were rated as achieving a light trance or above were symptom free 22 months later, none of the nonhypnotizable patients were improved. It was further pointed out that a strong relationship between the duration of the symptoms and the likelihood of therapeutic benefit existed. The likelihood of obtaining relief was markedly greater if therapy began shortly after the original trauma. The conclusion of Cedercreutz and his collaborators, that hypnosis is an appropriate treatment procedure for those patients with posttraumatic skull injuries who complain of headache and vertigo, does not necessarily follow, however. Earlier investigators, including Cedercreutz and Kampman (1972), found that in 3- to 4-year follow-ups on skull-injured untreated patients, 30%–34% still suffered from headaches (conversely 66%–70% spontaneously recovered). The present study, although it reported that 50% of the headache patients completely recovered and 20% partially recovered, did not take into account the base rate of spontaneous remission. Similarly, their finding that successful treatment was negatively correlated with the length of time between injury and the onset of treatment may be misleading. The symptoms of those patients who were treated long after their injury were evidently persistent and not likely to disappear over time, whereas patients who were seen immediately after an accident would be more likely candidates for spontaneous remission.

In a controlled study conducted by Andreychuk and Skriver (1975), biofeedback was compared to hypnosis as a treatment for migraine. In addition, the role of suggestibility in treatment was examined. Thirty-three established migraine sufferers were randomly placed in one of three groups: hypnosis treatment, hand warming training, or alpha enhancement training. All groups showed significant improvement from the baseline to the completion of 5 weeks of treatment. No treatment was found to be superior. Highly susceptible patients responded better than less susceptible ones, regardless of the type of treatment received. The initial goals of the biofeedback and hypnotic treatments differed, however. Biofeedback had as its intermediate goal the increase of blood flow and temperature in the hand, whereas the hypnosis treatment was geared exclusively to influence the headaches. It would have been valuable to set as an intermediate goal for both groups an increase in hand temperature and blood volume and to assess these goals prior to evaluating the effectiveness of each treatment in terms of symptom removal.

Anderson, Basker, and Dalton (1975) compared the effects of hypnotherapy and autohypnosis with a drug treatment program of Stemetil. The authors reported a difference in both number and intensity of attacks between the 6-month baseline and the first 6 months of hypnotic treatment. Stemetil treatment did not result in significant change. There was, however, no control for the added personal contact between physician and patient in the hypnosis treatment group, which could have accounted for the greater improvement obtained.

Insight. Blumenthal (1963) reported the successful treatment of two headache patients. The first patient had suffered with persistent pain over the left side of her neck and occiput. Through hypnosis, the patient recalled that her stepfather, who on several occasions had attempted to seduce her, had commonly placed his hand on her face and neck to comfort her and show her affection. In the second case, a Catholic father of five suffered from migraine headaches and complete sexual impotence. While hypnotized, he revealed that he was concerned about making his wife pregnant. Blumenthal hypothesized that the headaches served as a subconscious "displacement for his sexual feelings of erection and engorgement into the full engorged throbbing vascular headache" (p. 201). An exact report of the therapeutic procedures was not presented.

Altered perceptions. Kroger (1963) told headache patients, while they were hypnotized, to imagine their hand in ice water. They were then told that their hand was numb and anesthetized and that they could eliminate the pain in their head by placing this hand next to their head and face. Posthypnotic

suggestions that future headaches could be reduced in the same fashion were given. Kroger claimed that this procedure was successful but presented no data to corroborate the claim.

The case studies on headache fail to meet many of the previously mentioned methodological requirements. The treatment groups and controlled studies, in spite of the methodological problems mentioned, seemed to indicate that hypnosis may be a useful treatment adjunct for headaches. Again, as was found in the treatment of skin disorders, there were several interesting reports of using insight and altered perceptions to alleviate symptoms, but these too were methodologically unsophisticated.

Asthma

The third major psychosomatic disorder that was treated with hypnosis is asthma. Although there is no consensus as to the precise etiological factors involved in asthma, most authors agree that there is some interaction among psychological, infective, and allergic factors. Emotional factors are thought to be aggravating or triggering influences that induce the asthma attack in some patients (Peshkin, 1967). It is believed that anxiety that is associated with an anticipation of the symptoms themselves often results in an exacerbation of the symptoms (Hanley, 1974). Symptoms remain consistent even though causes may vary. During an attack, the air passages, including the trachea, major bronchi, and peripheral bronchioles, constrict. Bronchiospasms and increased secretion of mucus are also present. These physiological changes are responsible for the wheezing and difficulties in breathing seen in most asthmatic attacks.

Physiological changes. In an investigation by Dennis (1965), five good hypnotic subjects who had chronic asthmatic symptoms were exposed repeatedly to a true allergen and a placebo. Each exposure was given with suggestions for allergic reactions and no allergic reactions. This was done while the subject was either hypnotized or in the waking state. Patients were given a suggestion to experience a feeling of coolness. This presumably would induce vasoconstriction and thereby reduce asthmatic symptoms. Reactions to placebo could not be evoked in either the hypnotic or nonhypnotic conditions. However, allergic reaction to true allergen was successfully inhibited in the hypnotic condition. No attempt at monitoring the physiological mechanisms (i.e., vasoconstriction), which were assumed to mediate change in symptoms, was reported.

Weiss, Martin, and Riley (1970) found that suggestions of an asthma attack alone failed to produce allergic reactions in all but 1 of 16 asthmatic children, when they were presented with a saline solution. The 1 subject who reacted was thought to be allergic to the saline solution that was used in the bronchial challenge test.

Smith and Burns (1960) treated 25 chronic asthmatic children, ages 8–15. Hypnosis was achieved in all patients, and suggestions were given for symptomatic relief from asthma. Only objective measures of respiration were obtained, such as vital capacity and forced expiratory capacity. After four weekly sessions, no improvement was noted. These negative findings may be attributed to the unusually brief treatment and the sole use of physiological assessment.

Mun (1969) provided 10 asthmatic children with a series of treatments including one based on hypnotic suggestion. Each child was trained in hypnosis and was easily capable of entering into a trance. Each child, on the onset of an attack, was requested to report to the investigator for treatment. After the child reported, peak flow rate (a laboratory technique for assessing lung functioning) was measured, followed by one of four treatments. After 15 minutes of treatment, peak flow as well as self-report measures were obtained. Treatments consisted of Tedral medication after the first attack, Isoprenalinenebulizer during the second, hypnosis and hypnotic suggestions for symptom removal during the third attack, and Tedral medication and hypnotic suggestions during the fourth attack. Improvement was reported across all treatments, with hypnotic suggestion treatment showing the greatest effects, followed by the combined treatment of Tedral and hypnotic suggestion. No statistical analysis for any of the treatment comparisons was reported and treatment effects were obviously confounded.

Aronoff, Aronoff, and Peck (1975) reported on 17 asthmatic children who chose hypnotic treatment over traditional medical treatment. Hypnotic suggestions for symptom reduction were given during an actual asthmatic attack. Improvement was assessed in terms of physiological change and self-report. However, positive expectations for treatment and an expected decline in symptoms following the peak of an attack confounded the effects of hypnotic suggestion.

Philipp, Wilde, and Day (1972) looked at the effects of nonhypnotic suggestion on two types of asthma patients, those who reacted to skin testing (allergic) and those who did not (emotional). Both groups were subjected to two treatments, suggestions of a reaction when exposed to a neutral drug and suggestions of no reaction when an active drug was presented. Outcome measures included assessment of vital capacity and forced expiratory volume. It was found that emotionals reacted more to suggestions for symptoms on exposure to a neutral drug than did allergics. Another reported finding was that relaxation training across both groups tended to improve respiratory efficiency.

In a well-controlled study conducted by Fry, Mason, and Pearson (1964), the effects of hypnotic suggestion on allergic reactions to known allergens was assessed. Forty-seven asthmatic patients, who had displayed positive skin reactions to extracts of pollen or house dust and who were hypnotically susceptible, served as subjects. In the first part of the investigation, 18 subjects were randomly placed in either a control group, in which two skin tests were conducted, or in an hypnosis group, in which hypnotic suggestions for no skin reactions were given. Each subject was administered four solutions of varying strengths of an allergen. Results indicated that controls showed mixed reaction in terms of size of wheals after an exposure to allergens, whereas hypnosis subjects consistently showed a decrease in wheal size. In the second part of the investigation, 29 subjects were randomly divided among three groups: One received hypnotic suggestions that the right arm would not react to skin tests, another received suggestions that neither arm would react to testing, and the third received hypnosis with no suggestions. All groups showed marked decreases in wheal size. No differences were noted between the treatment groups, with hypnosis alone yielding as great a change as when it had been combined with suggestions.

Insight. Surprisingly, none of the asthma studies that were reviewed used hypnosis as a means of assisting patients in gaining insight into the maintenance of their symptoms.

Altered perceptions. The goal of the following investigations was to change the patients' perception of their symptoms so as to reduce the stress related to them and thereby reduce the intensity of the symptoms themselves.

Hanley (1974) described two cases in which hypnotherapy was used as a treatment for asthma. Therapy consisted of hypnotic suggestions to increase the individual's self-confidence, especially in controlling the asthmatic symptoms. Both patients reported some improvement in symptoms and in day-to-day living. It is not known, however, what effect the physician's supportive statements may have had without hypnosis.

Smith (1970) found, in two female subjects, an increase in pulmonary resistance following direct hypnotic suggestions of coughing, fear, anger, or of an imminent asthmatic attack and a decrease in resistance following suggestion of relaxation. Recall of previous asthma attacks could have been sufficient to create the change in pulmonary resistance.

Edwards (1960) used hypnotic suggestions for symptom reduction and a decrease in the psychological effects caused by an attack in six chronic asthma patients. Three assessment measures were employed: the patient's own testimony, stethoscope monitoring (physician's impression), and ventilatory function tests (vital capacity and forced expiratory volume). During and after treatment, patient reports revealed that their conditions improved. Physiological changes were not found, however. Although the author discussed the difficulty in making claims about treatment from a small uncontrolled study such as the one that he presented, this was not the most serious methodological flaw. Each of the cases were treated at the peak of an acute asthmatic attack. A regression

toward the mean would be expected on the assessment following the crisis, thus casting doubt on any claims of self-rated improvement due to treatment.

In a second investigation by Mun (1969), children were hypnotically age regressed to the time of their first attack and were told that the cause of the first attack was no longer operative and that they should no longer have fear and tension about the constantly recurring need to breathe. Mun reported improvement in self-ratings and objective measures, but, again, he presented no data to support these claims.

Moorefield (1971) described the successful treatment of nine patients with chronic asthma. Hypnotic procedures included suggestions for reduced tension and anxiety, for the ability to breathe easier, and for increased confidence during stress. Patients were seen for an average of 13.4 1-hour sessions. Complete recovery in all but one patient was reported. The effects of the hypnotic suggestions were confounded, however, since the patients were simultaneously treated with systematic desensitization.

The data on 121 asthmatic patients who were treated with hypnotherapy were analyzed retrospectively by Collison (1975). Treatment consisted of hypnotic suggestions that were intended to improve the patients' ability to cope with their environment and to assist them in obtaining both physical and mental relaxation. Posthypnotic suggestions for continued relaxation as well as training in autohypnosis were provided at each session. Patients were also classified according to the depth of trance typically achieved. Results indicated that 21% of the patients remained completely free of asthma attacks through the follow-up. Thirty-three percent showed improvement but continued to have mild attacks. Twenty-two percent improved, but symptoms were still moderately severe. Finally, 24% showed no changes from pre- to posttreatment. Patients who were rated as good hypnotic subjects received the most therapeutic benefit, whereas none of the patients who were poor hypnotic subjects became symptom free. Collison acknowledged that a retrospective analysis only allows for limited inferences. In addition, all patients had volunteered for hypnotherapy, which suggests that they may have had preexisting positive expectations for hypnosis. The care that Collison took in compiling his data was, nevertheless, commendable.

White (1961) hypnotically treated 10 asthmatic patients and obtained mixed results. Sessions consisted of hypnotic suggestions for easier breathing, lessening of tension and bronchiospasms, and an increase in self-confidence. Drug consumption and respiratory function as well as self-report were all assessed. Patients were rated in terms of depth of hypnosis achieved. Results varied depending on the particular measure used. Physiological measures indicated that improvement followed after 24% of the treatment sessions, whereas self-report of improvement followed 62% of the trials. Subject improvement was independent of the depth of hypnosis achieved.

Maher-Laughan (1970) reported three investigations that included autohypnosis training and hypnotic suggestions of a release in tension, an increase in self-confidence, and a belief in the possibility of recovery. In the first study, 55 patients were randomly placed in either a relaxation and breathing exercise control group or in a hypnosis and autohypnosis training treatment group. Change was assessed in terms of a computed wheezing score and in the number of times bronchodilators were used. Greater improvement was noted in the hypnosis treatment group. The second study included a larger number of patients who were treated by a greater number of centers and added a physiological assessment of respiratory function. Whereas males reported similar improvement in both the hypnosis group and the control group, females reported less improvement in the control group. The hypnosis group showed greater improvement on the physiological assessments of forced expiratory volume. Self-report and physiological indices were independent of the type of asthma treated (psychological, infectious, or allergic). The third study was similar to the first two except that patients were preselected for treatment, and a control group was not employed. In this investigation, 82% of the

patients improved. Duration of treatment was 6 months, 1 year, and 6 years, respectively, in the three studies, with few or no treatment effects recognized earlier than 1 month into treatment.

The greatest number of controlled studies on the largest number of subjects has been carried out with asthmatic patients. The methodological sophistication of these studies far exceeds the studies of skin disorders or headaches. Effectiveness of hypnotic treatment appears to be a function of which outcome measures are employed. When self-report is used, most investigators report an improvement in the patient's symptoms. However, physiological measures show more equivocal results.

Discussion and Conclusions

Of the 38 studies reviewed, 31(81.6%) reported overall positive results, 5(13.1%) reported negative outcomes, and 2(5.3%) reported mixed results. Eighteen (47.4%) were single- or multiple-case studies, 13(34.2%) used a single treatment group, and 7(18.4%) instituted necessary control conditions. All of the single- and multiple-case studies resulted in a positive outcome. Ten single treatment group studies yielded a positive outcome, 1 yielded a negative result, and 2 yielded a mixed outcome. Three of the 7 controlled studies resulted in a negative outcome. Apparently when methodological shortcomings were remediated, the effects of hypnosis were reduced.

There were 1,189 subjects/patients in the 37 studies that reported sample size. (Kroger, 1963, did not report the number of patients treated.) Single- and multiple-case studies involved only 35 patients (2.9% of the total) even though case studies constituted almost half of the studies reviewed. Studies obtaining positive results involved 1,037(87.2%) subjects/patients. Only 152 (12.8%) subjects/patients were involved in studies that yielded negative or mixed results.

A comparison of the three disorders shows that 67.3% of the patients were asthmatic, 23.4% were headache patients, and 9.2% had skin disorders. Inspection of Table 1 reveals that hypnosis was most successful for asthmatics; all but two studies in Table 1 used at least a treatment group. Results were positive in 11, negative in 2, and mixed in 2 studies. For headaches, 5 out of 9 were case studies, and 1 of the only 2 controlled studies yielded negative results. Only 3 out of the 14 studies of skin disorders used at least a treatment group. The 2 controlled studies yielded negative results.

Changes in autonomic nervous system functioning (i.e., blood flow, skin temperature, forced air capacity) are thought to serve as mediators of symptomatic changes. Twenty-three (60.5%) studies used hypnosis in an attempt to directly affect physiological change (i.e., the blood flow to the skin will decrease). In only 8 of these 23 studies was there an attempt to assess these changes. Four of the 8 reported negative results.

Four (10.5%) of the studies used hypnosis to help the patient gain insight. All reported success, but this was based on self-reports or physicians' impressions, and no objective measure of success was reported.

Of 10(26.3%) studies, in which hypnosis was used to alter patients' perceptions of their disorders, 8 reported successful outcomes, and 2 reported mixed results.

Physiological or objective indices were used in only two studies of skin disorders, with negative results. Understandably physiological or objective measures were not used in the headache studies because headache pain is almost exclusively a subjective experience. Vasoconstriction and muscular tension have been found to correlate with headache pain and could have been used. In the asthma studies, 11 used a physiological or objective measure. Positive outcomes were reported in 7, and negative results were reported in 4 studies. A much higher rate of success was found for the 32 studies using self-report as a measure. All outcomes but 1 were positive.

It appears the attempts to alleviate symptoms by suggesting direct and specific physiological changes through hypnosis has met with some success, as measured by self-report, but results that used objective or autonomic measures have been equivocal.

None of the investigations of headache attempted to assess muscular or vascular changes. The studies of skin disorders as-

sessed the physical changes in the skin. However, mediating physiological changes such as blood flow and skin temperature were generally not monitored. The change in symptoms (disappearance of warts, changes in pruritus, etc.) cannot be taken as evidence for direct physiological change due to hypnosis. Controls were not sufficiently applied to show a cause and effect relationship between the applications of the hypnotic procedures and the alleviation of symptoms. Evidence from the asthma studies supports the contention that hypnosis does not produce physiological changes as consistently as it produces self-reported and observed changes. This is in agreement with arguments presented by Barber (1974) who contends that "the data available at present do not support the notion that hypnotic trance is a critical factor in producing such physiological effects" (p. 70).

There seems to be some anecdotal evidence that alleviation of symptoms may be aided by using hypnosis to help patients to gain insight into the development and maintenance of their symptoms. A difficulty with these investigations is a lack of an established cause and effect relationship between insight and symptom remission. The possibility of spontaneous remission or placebo effects was not controlled for in these case studies.

Evidence exists for the claim that when hypnosis is used to alter the perceptions of psychosomatic symptoms, headache and asthma patients report symptom improvement. This appears to be especially useful either in disorders in which the perception of the symptoms themselves is the primary concern (i.e., tension headaches) or when the disorder is being treated by an additional treatment regime and symptomatic relief is desirable.

Future researchers, in addition to stating the ultimate goal of the hypnotic intervention, also should clearly specify the mediating physiological mechanisms for the expected change. The precise role of these mechanisms could then be systematically assessed.

Controls should be instituted in this research for suggestions alone, hypnosis alone, and physician and patient interaction. In this way, the role of these variables could be separately assessed. In addition, the possibilities of spontaneous remission of symptoms and placebo effects, which seem to occur with great frequency in psychosomatic disorders, should be controlled.

Researchers should assess hypnotic susceptibility using a well-established scale. Although some researchers have argued for the placement of subjects into groups based on their susceptibility, this procedure would limit the extent to which a cause–effect relationship could be inferred between treatment and outcome. A more useful procedure would be to randomly place subjects, regardless of their susceptibility scores, into treatment groups while simultaneously monitoring the effects of susceptibility on outcome.

In spite of the many methodological problems, there was some indication of the usefulness of hypnosis as a treatment adjunct for psychosomatic disorders. This is particularly true for asthma, the disorder on which the best research has been done and for which the most promising results obtained.

References

Ament, P., & Milgrom, H. Effects of suggestion on pruritus with cutaneous lesions in chronic myelogenous leukemia. *Journal of the American Society of Psychosomatic Dentistry and Medicine,* 1967, *14,* 122–125.

Anderson, J. A., Basker, M. A., & Dalton, R. Migraine and hypnotherapy. *International Journal of Clinical & Experimental Hypnosis,* 1975, *23,* 48–58.

Andreychuk, T., & Skriver, C. Hypnosis and biofeedback in the treatment of migraine headache. *International Journal of Clinical & Experimental Hypnosis,* 1975, *23,* 172–183.

Ansel, E. L. A simple exercise to enhance response to hypnotherapy for migraine headache. *International Journal of Clinical & Experimental Hypnosis,* 1977, *24,* 68–71.

Aronoff, G. M., Aronoff, S., & Peck, L. W. Hypnotherapy in the treatment of bronchial asthma. *Annual of Allergy,* 1975, *34,* 356–362.

Barber, T. H., Spanos, N. P., & Chaves, J. F. *Hypnotism: Imagination and human potentialities.* New York: Pergamon Press, 1974.

Blumenthal, L. S. Hypnotherapy of headache. *Headache,* 1963, *2,* 197–202.

Cedercreutz, C., & Kampman, R. Hypnotic treatment of posttraumatic headache. *Hypnosis and Psychosomatic Medicine,* 1972, *21,* 150–151.

Cedercreutz, C., Lahteenmaki, R., & Tulikoura, J. Hypnotic treatment of headache and vertigo in

skull injured patients. *International Journal of Clinical & Experimental Hypnosis,* 1976, *24,* 195–200.

Clawson, T. A., & Swade, R. H. The hypnotic control of blood flow and pain: The cure of warts and the potential for the use of hypnosis in the treatment of cancer. *American Journal of Clinical Hypnosis,* 1975, *17,* 3, 160–169.

Collison, D. R. Which asthmatic patients should be treated by hypnotherapy? *Medical Journal of Australia,* 1975, *1,* 776–781.

Dennis, M. Hypnotic and non-hypnotic suggestion and skin response in atopic patients. *American Journal of Clinical Hypnosis,* 1965, *7,* 342–345.

Edwards, G. Hypnotic treatment of asthma: Real and illusory results. *British Medical Journal,* 1960, *5197,* 492–497.

Ewin, D. M. Condyloma Acuminatum: Successful treatment of four cases by hypnosis. *American Journal of Clinical Hypnosis,* 1974, *17,* 73–78.

Frankel, F. H., & Misch, R. C. Hypnosis in a case of long-standing psoriasis in a person with character problems. *International Journal of Clinical & Experimental Hypnosis,* 1973, *2,* 121–130.

French, A. P. Treatment of warts by hypnosis. *American Journal of Obstetrics and Gynecology,* 1973, *116,* 887–888.

Fromm, E. Altered state of consciousness and hypnosis: A discussion. *International Journal of Clinical & Experimental Hypnosis,* 1977, *25,* 325–334.

Fry, L., Mason, A. A., & Pearson, R. S. Effects of hypnosis on allergic skin responses in asthma and hay fever. *British Medical Journal,* 1964, *5391,* 1145–1148.

Graham, G. W. Hypnotic treatment for migraine headaches. *International Journal of Clinical & Experimental Hypnosis,* 1975, *23,* 164–171.

Hanley, F. W. Individualized hypnotherapy of asthma. *American Journal of Clinical Hypnosis,* 1974, *16,* 275–279.

Harding, H. C. Hypnosis and migraine or vice versa. *Northwest Medicine,* 1961, *60,* 168–172.

Hilgard, E. R. *The experience of hypnosis.* New York: Harcourt, Brace & World, 1968.

Jabush, M. A case of chronic recurring multiple boils treated with hypnotherapy. *Psychiatric Quarterly,* 1969, *43,* 448–455.

Johnson, R. R., & Barber, T. X. Hypnotic suggestions for blister formation: Subjective and physiological effects. *American Journal of Clinical Hypnosis,* 1976, *18,* 172–181.

Klinge, J. E. Atopic dermatitis. *Journal of the American Institute of Hypnosis,* 1971, *12,* 128–131.

Kroger, W. S. Hypnotherapeutic management of headache. *Headache,* 1963, *3,* 50–62.

Maher-Laughan, G. P. Hypnosis and autohypnosis for the treatment of asthma. *International Journal of Clinical & Experimental Hypnosis,* 1970, *1,* 1–14.

Moorefield, C. The use of hypnosis and behavior therapy in asthma. *American Journal of Clinical Hypnosis,* 1971, *13,* 162–168.

Mun, C. T. The value of hypnotherapy as an adjunct in the treatment of bronchial asthma. *Singapore Medical Journal,* 1969, *10,* 182–186.

Peshkin, M. Incidence and etiology of asthma in childhood. *Journal of Asthma Research,* 1967, *4,* 179–182.

Peters, J. E., & Stern, R. M. Specificity of attitude hypothesis in psychosomatic medicine: A reexamination. *Journal of Psychosomatic Research,* 1971, *15,* 129–135.

Philipp, R. L., Wilde, G. S., & Day, J. H. Suggestions and relaxation in asthmatics. *Journal of Psychosomatic Research,* 1972, *16,* 193–204.

Smith, J. M. Increase and decrease in pulmonary resistance with hypnotic suggestion in asthma. *American Review of Respiratory Disease,* 1970, *102,* 236–242.

Smith, J. M., & Burns, C. L. The treatment of asthmatic children by hypnotic suggestion. *British Journal of Diseases of the Chest,* 1960, *54,* 78–81.

Surman, O. S., Gottlieb, S. K., Hackett, T. P., & Silverberg, E. L. Hypnosis in the treatment of warts. *Archives of General Psychiatry,* 1973, *28,* 439–441.

Tasini, M. F., & Hackett, T. P. Hypnosis in the treatment of warts in immunodeficient children. *American Journal of Clinical Hypnosis,* 1977, *19,* 152–154.

Todd, F., & Kelly, R. The use of hypnosis to facilitate conditioned relaxation responses: A report of three cases. *Journal of Behavior Therapy and Experimental Psychiatry,* 1970, *4,* 295–298.

Twerski, A. J., & Naar, R. Hypnotherapy in a case of refractory dermatitis. *American Journal of Clinical Hypnosis,* 1974, *16,* 202–205.

Vollmer, H. Treatment of warts by suggestion. *Psychosomatic Medicine,* 1946, *8,* 138–142.

Weiss, J. H., Martin, C., & Riley, J. Effects of suggestion on respiration in asthmatic children. *Psychosomatic Medicine,* 1970, *32,* 409–415.

White, H. Hypnosis in bronchial asthma. *Journal of Psychosomatic Research,* 1961, *5,* 272–279.

Wolff, H. G. *Headache and other head pain.* New York: Oxford University Press, 1963.

Received June 5, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1236–1254

Stimulus Overselectivity in Autism: A Review of Research

O. Ivar Lovaas
University of California, Los Angeles

Robert L. Koegel
University of California, Santa Barbara

Laura Schreibman
Claremont Men's College

Infantile autism is a severe form of psychopathology characterized by profound behavioral deficits. This article reviews a series of investigations which suggest that autistic children show "stimulus overselectivity," a response to only a limited number of cues in their environment, and discusses how such overselectivity may relate to several of the behavioral deficits in autism. These include failure to develop normal language or social behavior, failure to generalize newly acquired behavior to new stimulus situations, failure to learn from traditional teaching techniques that use prompts, and a general difficulty in learning new behaviors. This discussion is followed by the presentation of several studies that suggest possible remedial procedures. Finally, the concept of stimulus overselectivity is related to the literature on other theories of attentional or response deficits in adult schizophrenia, mental retardation, learning disabilities, and autism.

Infantile autism, first described by Kanner (1943), is a severe form of psychopathology in children that is characterized by extreme social and emotional detachment. Such children typically do not seek or readily accept affection and do not play with peers. They engage in great amounts of stereotyped, ritualistic, and repetitive motor behaviors and are generally unresponsive to their physical environment. They are inconsistent in their response to sensory input, they typically do not show a startle reflex, and their parents have suspected them to be blind or deaf. Language development is either absent or abnormal, and those autistic children who do speak usually parrot meaninglessly what they hear (echolalia). Another characteristic feature of autism is the child's insistence on order and sameness in his or her environment. When one considers the behavioral impoverishment of these children, it is understandable that autism is also characterized by a poor prognosis.

In recent years, numerous studies have appeared that somewhat alter the prediction of a poor prognosis. For example, autistic children who are treated within a learning theory framework (with "behavior modification") have shown measurable improvement in speech and language (e.g., Hewett, 1965; Lovaas, 1966, 1977; Risley & Wolf, 1967), generalized imitation (Lovaas, Freitas, Nelson, & Whelan, 1967; Metz, 1965), and appropriate play (Koegel, Firestone, Kramme, & Dunlap, 1974), as well as reduction of inappropriate behaviors (e.g., Carr, Newsom, & Binkoff, 1976). A comprehensive summary of these and other studies has been provided by Lovaas and Newsom (1976). However, in spite of considerable successes, there are still a number of problems associated with this

The preparation of this article and portions of the research reviewed were funded by U.S. Public Health Service Research Grants MH 11440, MH 28210, and MH 28231 from the National Institute of Mental Health and by U.S. Office of Education Research Grant G007802084 from the Bureau for the Education of the Handicapped.

The authors wish to gratefully acknowledge the administration and staff of Camarillo State Hospital, Camarillo, California, for their assistance and cooperation in the collection of much of the data reported here.

Requests for reprints should be sent to O. Ivar Lovaas, Department of Psychology, University of California, Los Angeles, California 90024.

form of intervention. Two important problems are that the treatment effects may be situation specific and reversible, and the child's progress in treatment is slow (Lovaas, Koegel, Simmons, & Long, 1973).

Problems in situation specificity of the treatment change and in reversibility of treatment effects can be attenuated somewhat by extending the treatment across time and situations, by, for example, using parents and teachers as therapists (e.g., Schreibman & Koegel, 1975). However, there has been no immediate solution to the slow rate of improvement that these children show in treatment. It may be most helpful, therefore, to discuss certain mechanisms that may be basic to this problem of attenuated change, with a view toward the development of faster and more general changes. Towards this end, this article will review a set of studies that provide substantial evidence which suggests that autistic children have a problem in responding to multiple cues, and that this problem may be at least part of the cause of their slow improvement in treatment.

Situations in which several cues impinge simultaneously on children are typical of those they encounter in their everyday teaching environments. Operationally, our data show that when autistic children are presented with multiple stimulus inputs, their behavior comes under the control of a range of input that is too restricted. This problem was referred to as "stimulus overselectivity" (Lovaas, Schreibman, Koegel, & Rehm, 1971) because the children overselected a limited set of stimuli from those available in their environment. Note that the term *stimulus overselectivity* does not imply that the children scan their environment and select for relevant cues. Rather, the data suggest that the children respond to only part of a relevant cue, or even to a minor, often irrelevant feature of the environment, without learning about the other relevant portions of that environment.

Our approach to this analysis in autistic children owes much to the conceptual and methodological advances of basic operant research, as first reviewed by Terrace (1966) and later extended by Ray (1972), Ray and Sidman (1970), Sidman and Stoddard (1966), and Touchette (1968, 1971). Basically, the studies we report adhere to a discrimination learning paradigm, permitting us to simply relate aspects of the manipulated stimulus input directly to the child's behavior. Excellent reviews of this type of research have also been provided by Fellows (1968), Sutherland and MacKintosh (1971), and Trabasso and Bower (1968).

This article first reviews a set of studies demonstrating stimulus overselectivity, and then presents data that show how such overselectivity may interfere with the autistic child's learning and pose problems for the generalization of learned material. This will be followed by suggestions on how this problem might be remedied. Finally, the concept of stimulus overselectivity will be related to other theories of attentional deficits in autism and related disorders.

Studies Demonstrating Stimulus Overselectivity

Overselective Response to Multiple Cues

In the Lovaas et al. (1971) study, autistic, retarded, and normal children were taught to respond to a complex stimulus display (S^D) containing three elements: (a) a moderately bright visual stimulus (consisting of a 160-W red floodlight), (b) an auditory stimulus consisting of white noise at a moderately high (65-dB level) intensity, and (c) a tactile stimulus on the child's leg delivered by a pressure cuff at 20 mm of mercury. These stimuli appeared noticeable to the children, since they often oriented to them (e.g., turned around to look at the light, touched their legs when the cuff was inflated). This complex S^D was presented to the children, and they were reinforced for responding (bar pressing) in the presence of the display and not reinforced for responding in its absence. After training had established this stimulus display as functional for the children's response, single-cue test trials were presented in which each component (auditory, visual, tactile) was presented separately for a total of 70 presentations over 10 test sessions.

The results showed that the normal children responded to each of the components equally. In other words, each of the separate cues became equally functional in controlling the

child's behavior. The performance of the autistic children was different. Each child responded primarily to only *one* of the component cues. (The retarded children responded at a level between these two extremes.) Three of the autistic children responded primarily to the auditory component, whereas two of the children responded primarily to the visual cue. None of the autistic children responded to the tactile stimulus. It was striking to observe the autistic children attentively respond to one of the component cues (e.g., the sound), only to remain motionless in the presence of the other (e.g., the light) even though that stimulus had been presented as discriminative for reinforcement.

Subsequently, two of the autistic children were trained to respond to the component that had remained least functional for them during test sessions. Both children quickly learned to respond to the previously nonfunctional component, *when that component was presented alone.* This helped to ensure that the problem was not one of some relatively "simple" sensory deficit (as in the case of being blind or deaf) but was rather a problem in responding to the cue in the context of other cues.

It was concluded from this first study that the data could best be understood as representing the autistic child's difficulty in responding to stimuli in context, a problem pertaining to the quantity rather than to the quality of stimulus control. The data failed to support any notion that a particular sense modality was impaired in autistic children or that any particular sense modality was a "preferred" modality.

Overselective Response to Two Cues

Since the autistic child may have been "flooded" or "overloaded" with stimulation in the first study, a second study (Lovaas & Schreibman, 1971) was conducted in which the stimulus input was simplified: The child was presented with only two cues. The two cues were the same red floodlight and white noise used in the previous study. The experimental paradigm was also the same as in the previous study.

It may be of interest to examine the data from this study in some detail, since they show certain unpredictable peculiarities. The six normal children tested gave no signs of stimulus overselectivity; they responded equally to the two components when these were presented on separate occasions during test trials. It was different with the autistic children. The data are presented in Figure 1. Four of the children showed the stimulus overselectivity most clearly. For each of these children only one of the cues was functional in controlling their behavior. Two of these children (Kevin and Michael) displayed some response to both component cues during early parts of the testing but eventually lost rather than acquired the response to the nondominant cue. The fifth and sixth children (Janet and John T.) showed stimulus overselectivity in the early test trials but gradually began responding to the initially weak component cue as testing progressed. They may have overcome their overselectivity with testing. Another child, Jimmy, showed control by one of the stimuli at first (the visual one); but as testing progressed, for some reason the visual stimulus lost its control, whereas the auditory stimulus gradually assumed control. Bobby showed only a slight case of stimulus overselectivity, if any. Finally, the last child, John P., responded equally to both components and showed no evidence of stimulus overselectivity.

Thus stimulus overselectivity was not observed for all the autistic children in this study, whereas all of the children gave evidence of stimulus overselectivity in the previous study, which involved three cues. It may be that stimulus overselectivity is most clearly observed with a relatively larger quantity of stimulus inputs.

Discrimination of a Complex Stimulus From Its Components

Several questions can be raised as to why the autistic children responded the way they did on these tests. One possibility is that autistic children have a genuine difficulty in responding to the separate components of a complex input. Another possibility is that autistic children do respond to the components

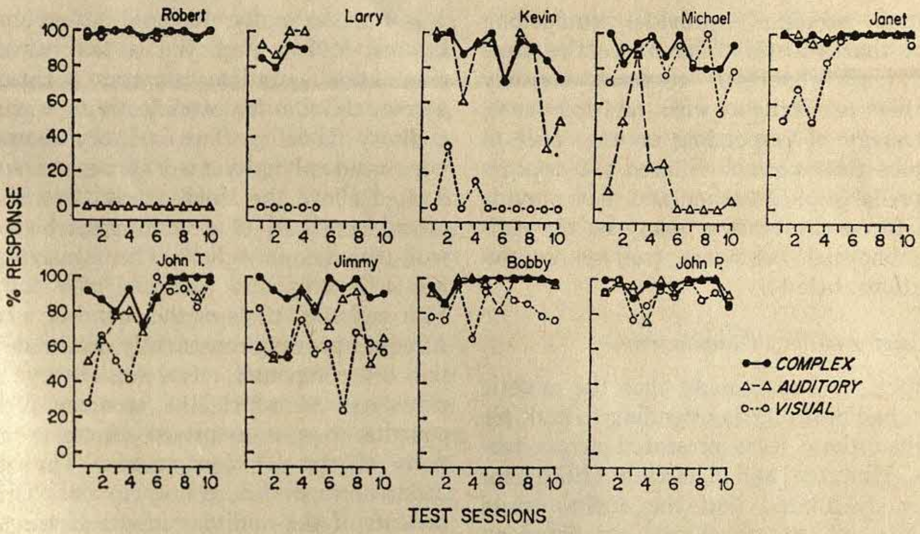

Figure 1. Test sessions for the autistic subjects. (Percentages of correct responses to stimuli are plotted on the ordinate, and test sessions are plotted on the abscissa.)

but that they are "superefficient," so that they immediately know (discriminate) that it is enough to respond to only one component of a complex input to be reinforced. They may exert minimum effort for maximum payoff.

Koegel and Schreibman (1977) conducted a study to help answer these questions. The procedures and apparatus were similar to those employed in the first two studies. Subjects were first trained to respond to the separate presentations of an auditory and then a visual stimulus. That is, the components were presented first. At the completion of the training phase the child was presented with either the separate cues or these cues combined in a complex. Thus, three types of trials were presented: visual only, auditory only, and visual and auditory combined. All responses to the complex auditory–visual input were reinforced, whereas responses to either of the single cues were not reinforced.

The most general conclusion to be drawn from the results of this study is that in this conditional discrimination task, autistic children experienced great difficulties discriminating the complex stimulus from its components. This conclusion is based on the fact that the autistic children continued to respond to *one* of the components, for hundreds of trials, even though they were not reinforced for doing so; concurrently, their response to the other (also nonreinforced) cue extinguished relatively rapidly. In contrast, the normal children extinguished on both of the single nonreinforced components quickly and simultaneously. Since the autistic children continued to respond to one of the single components, even though such responding was not reinforced, it does not seem likely that stimulus overselectivity shown by autistic children is based on some variable such as efficient responding. To respond consistently to one, and only one, of the nonreinforced components as well as to the reinforced stimulus complex seems like a fairly complicated strategy, unless one postulates that the complex and one of the components were difficult to discriminate from each other. It may also be important to note that the autistic children did eventually extinguish responding to both component cues. Thus, it appears that the autistic children were attempting to learn the discrimination but had difficulty doing so.

Another major point of this experiment is that with a completely different experimental paradigm, the same type of result was obtained as in the other studies. That is, the consistency in the results across all of these experiments (regardless of the specific ex-

perimental paradigm) provides compelling evidence that autistic children use the same specific (and abnormal) response strategy when they are taught with multiple cues. This strategy of responding on the basis of fewer cues than normal children use appears to be a reliable phenomenon and may provide a basis for understanding many of the children's abnormal behaviors (see section on Implications, below).

Visual and Auditory Overselectivity

Initially, we had thought that the autistic children had difficulty in attending to multiple cues when these were presented across modalities. Hintgten and Churchill (1971), for example, speculated that the autistic child may have an attentional problem based on some difficulty with "integrating" information along more than one modality. Subsequent studies in our laboratory suggest that autistic children have difficulty with multiple cues even when these cues are presented within the same modality.

Koegel and Wilhelm (1973) trained 15 autistic children and 15 normal children of approximately the same chronological age on a visual discrimination task. Their procedure followed the relevant redundant cue paradigm. The children were trained to discriminate between two cards, each card containing two visual cues. After the children had mastered this discrimination (consistently responding to one of the cards), test trials were conducted in which the visual cues on the cards were split so that one of the components of the correct card was presented versus one of the components on the incorrect card. The autistic children gave evidence of overselective responding in this situation by reliably choosing a card with one of the components of the original complex correct stimulus and responding at chance level to the other component. The normal children, on the other hand, responded predominantly to both the components during testing. At an even finer level of analysis, Koegel and Schreibman (1977) found that low-functioning autistic children show visual overselectivity even between the components of a single visual stimulus, such as selectively responding to the color, form, or shape of a triangle.

In another study, Reynolds, Newsom, and Lovaas (1974) employed a successive discrimination paradigm to test autistic and normal children for overselectivity within the auditory modality. Two auditory compounds were randomly presented through a speaker located above the child's head. The S^D compound consisted of a continuous high tone, with periodic relay clicks. The negative stimulus (S^Δ) compound consisted of a low tone with periodic bursts of the sound of a motor. After bar pressing consistently occurred during the S^D compound, test sessions were administered in which the separate S^D components were interspersed among presentations of the S^Δ components. The normal children responded reliably to both S^D components of the auditory input. However, the autistic children responded to one S^D component, giving evidence for stimulus overselectivity within the auditory modality.

The studies reviewed above are consistent with studies from other laboratories which suggest that autistic children have particular difficulties when they are expected to associate multiple stimuli. For example, Cowan, Hoddinott, and Wright (1965) found that only 2 of 12 autistic children were able to associate simple shape or color words with corresponding visual stimuli. Bryson (1970) noted that autistic children performing match-to-sample tasks disregarded extra visual stimuli when they made vocal responses. Similarly, Frith and Hermelin (1969) found that if normal children were blindfolded during maze training tasks, they were relatively more handicapped than when autistic children were blindfolded. Apparently, the autistic children were less affected by additional cues than were the normal children.

Stimulus Overselectivity, IQ, and Mental and Chronological Age

Although the preceding studies suggest that autistic children show stimulus overselectivity, one must remain skeptical about the role of such overselectivity in the etiology of autistic behavior for at least two reasons. First, since these studies provide correlational data only, stimulus overselectivity could just as well be the effect of autistic behavior as its cause.

Second, in the studies we have described so far, a few autistic children showed little or no evidence of overselectivity, whereas some children who were not autistic did overselect. For example, in the Lovaas et al. (1971) study discussed previously, retarded (but nonautistic) children did overselect.

A relationship between IQ level and stimulus overselectivity was demonstrated in a study by Wilhelm and Lovaas (1976) that reported on three groups of children with different IQ levels. On a discrimination task that could be solved by the child attending to either one, two, or all of three component cues, the low IQ (20) group responded on the average to 1.6 cues, the higher IQ (40) group responded to 2.1 cues, and the normal IQ children responded to all three cues.

A similar finding has also been reported on the relationship between chronological age and stimulus overselectivity. Schover and Newsom (1976) found that the younger the children, the more likely they were to show overselectivity. The Schover and Newsom findings are in accord with other studies that show a positive correlation between number of cues responded to in the visual discrimination task and mental age in normal and retarded children (Eimas, 1969; Fischer & Zeaman, 1973; Hale & Morgan, 1973; Olson, 1971).

It remains to be determined whether this general finding that relates mental age to overselectivity will hold up with multimodal stimuli, as in the first studies presented here (Lovaas & Schreibman, 1971; Lovaas et al., 1971). Until studies employing such stimulus compounds are carried out, the status of overselectivity as an etiological factor in autism remains in doubt; as Schover and Newsom (1976) and Sivertsen (1976) have pointed out, young normals showing visual overselectivity seem intuitively to be far ahead of their autistic counterparts in linguistic, intellectual, affective, and social behaviors. Nevertheless, autistic or retarded children who remain overselective long after their normal peers have moved beyond this mode of functioning are clearly at a disadvantage in acquiring new behavior, and overselectivity undoubtedly contributes to the maintenance of such children's behavioral retardation (Ross, 1976; Wilhelm & Lovaas, 1976). That is, the more limited the aspects of the environment become in controlling the children's behavior, the more retarded will be their behavioral development. This conclusion will become apparent when we discuss the relationship of overselectivity to learning situations that require shifts in stimulus control and the generalization of acquired behaviors across environments.

Implications

Some of the implications of these findings for understanding autistic development seem obvious and were pointed out in an earlier article (Lovaas et al., 1971). We argued that many learning situations in life necessitate responding to multiple cues. Speech, for example, is a complex stimulus input for which adequate responding necessitates the child's attention to a number of stimulus dimensions (e.g., voiced vs. voiceless, tense vs. lax, volume vs. pitch). If a child responds to only one or two of these dimensions, he or she will not understand what is said. Reynolds et al. (1974) speculated that autistic children's deficiency in language may be based on their failure to respond to multiple inputs. Similarly, much of the meaning in language involves associating multiple inputs, as in associating language cues to a variety of sensory inputs such as sight and feel. We also speculated that stimulus overselectivity may underlie autistic children's deficiency in emotional behavior. Thus we pointed out that classical conditioning is considered by many to underlie the acquisition of emotional behavior. The latter requires attention to two or more contiguous or nearly contiguous stimuli (the conditioned and the unconditioned stimulus). In such a situation autistic children may well overselect, responding to one or the other of these stimuli, but not both. If they fail to respond to both, they may fail to condition (Maltzman & Raskin, 1965).

In the following section we try to support with data some of these speculations on how stimulus overselectivity interferes with autistic children's learning and with the generalization of that learning to new environments.

Observational Learning

Undoubtedly, many of the complex and subtle behaviors that are shown by normal children are learned by watching social interactions of various kinds (Bandura, 1969). Children who do not naturally learn in this way can be expected to show considerable behavioral retardation. Observational learning seems to be an area in which autistic children are particularly handicapped, and their failure to learn through observation may well be basic to one's understanding of their behavioral deficiencies (Ross, 1976). Typically, observational learning necessitates attention to multiple cues, in observing both a model's behavior and the consequences of that behavior.

Varni, Lovaas, Koegel, and Everett (1979) obtained data which suggest that stimulus overselectivity may prevent observational learning in autistic children. In this study, the autistic children sat at a table across from two adults (a model and a teacher). On the table in front of them rested two objects, each of which was used in the execution of an associated response. The investigators assessed the extent to which the children could learn how to behave in this situation by merely observing the model handle the objects in accordance with the teacher's instructions. The teacher gave the model a command (e.g., "phone"), and the model then behaved accordingly (e.g., picked up the handset of the phone) and was rewarded by the teacher. A sequence of 20 observation trials was followed by one test trial to see if the children had acquired the task. During the test trials the children took the model's seat and were given the same command by the teacher. Correct performance on a test trial was taken to indicate that observational learning had occurred, since the children were not directly taught by the teacher to respond correctly. The observation trials continued until the children responded correctly on a test trial, or until they had experienced 1,000 trials without showing evidence of learning. If this first task was mastered, a second response that involved a second object was presented (e.g., placing a ball in a toy dump truck when the teacher said "dump"); and if tests showed that the children mastered both responses, they were given 10 additional test trials with the two verbal commands presented in random order.

Data from this study showed that the autistic children usually learned only part of the response they observed. For example, a child might touch rather than pick up the phone or merely touch or move the dump truck. In some cases the children learned the complete behavioral topography but did not associate it with the teacher's verbal command. For example, a child might pick up the phone regardless of whether the teacher said "phone" or "dump." In summary, the children's failures could be related to their responses to only restricted portions of the complex stimulus situation that they had observed.

Prompt Studies

Another kind of learning in which stimulus overselectivity seems particularly handicapping is in the area of prompting and prompt fading. Prompts are usually extra stimuli added to the learning situation to ensure correct responding. In this kind of learning a teacher may help the children associate a response to a particular stimulus by first prompting the right response. Many learning situations involve such prompts or "guidance" because the teacher may not be able to wait for the children to give the correct response on their own. For example, in teaching children to read a new word, the teacher may prompt them by also presenting a picture of the referent. If such learning is to be successful, such prompting or guidance must eventually be removed or "faded" so that the children behave on their own. Technically, such learning is referred to as the acquisition of *stimulus control*, and involves shifts in stimulus control from prompt stimuli (e.g., picture) to training stimuli (e.g., written word).

The stimulus overselectivity hypothesis suggests that the provision of additional stimuli in prompt procedures may prevent the children from learning. That is, most prompt procedures require the children to respond to multiple cues (the prompt and training stimuli occur together) and should create situations in which stimulus overselectivity is likely to occur.

Koegel and Rincover (1976) provided data that show the deleterious effects on learning when extra cues are used as prompts. They pretrained autistic and normal subjects to respond differentially to two easily discriminable colors (red and green). Once this discrimination was mastered, the colors were used as prompts and presented simultaneously with more difficult discriminations. (e.g., A low-pitched tone was presented concurrently with the color red and a high-pitched tone with the color green.) The colors were then eliminated gradually, the auditory stimuli alone remaining. The normal children learned the new (e.g., tone) discrimination this way, whereas the autistic children usually failed to transfer from the prompt to the training stimuli. Instead they continued to respond to the color cues even when they were faded to a barely recognizable level. However, the autistic children did acquire these new discriminations when prompts were not used in training. The Koegel and Rincover data suggest that the use of extra cues may make it more difficult for the autistic children to learn. Schreibman (1975) observed the same problem. The autistic children responded selectively to the prompt as long as it was available but reverted to chance performance when the prompt was removed.

Such problems with prompts may happen more often when the discrimination involves difficult rather than easy discriminations, as shown in a study by Russo (Note 1). He reported that autistic children shift from a prompt (finger pointing) to training stimuli more readily when the task is easy (e.g., black vs. white) than when it is difficult (e.g., vertical line vs. a slightly tilted line).

Generalization Studies

Restrictions on the number of stimuli that acquire control over behavior could cause serious problems in stimulus generalization, that is, the extent to which a behavior learned in one environment transfers to other new environments. This relates to the familiar problem of "undergeneralization" of therapeutic gains—the failure of a behavior, acquired in a therapeutic setting, to transfer to a new "outside" environment.

One can consider generalization to take place to the extent that there are common stimulus elements between the teaching situation and outside situations. Amount of generalization, then, may vary proportionately with the number of stimulus elements that controlled the behavior initially. The fewer the stimuli that had become functional in the original situation, the fewer stimulus elements that would control the behavior in the new environment; hence limited generalization occurs. This latter problem, that of limited generalization, has been clinically observed in all our work with autistic children. It is shown in a study by Rincover and Koegel (1975), in which stimulus overselectivity seemed to directly limit generalization. In this experiment, one teacher taught autistic children to perform a simple behavior on request (e.g., "touch your nose"). Immediately after each child had learned this behavior, a second teacher took the child into another environment and made the same request. Four of the 10 autistic children did not perform the relatively simple behavior in the new environment. Extensive single-subject analyses showed that those four children had failed to generalize the learned behavior because they had selectively responded to irrelevant stimuli during the original training and had not learned the response on the basis of the relevant cue. In one case, for example, the child's responding was controlled by incidental movements of the teacher's hand, and not by the relevant verbal cue. (i.e., Without the incidental hand movement the child would not respond to the verbal cue, whereas the child would respond to the hand movement either alone or with the verbal cue.)

Note that this result is different from saying that the autistic children responded to too many cues (i.e., central plus incidental cues; cf. Tarver, Hallahan, Kauffman, & Ball, 1976). Rather, the autistic children responded to too few cues so that in some instances, they responded only on the basis of an incidental cue and not on the basis of the central cue; thus it appeared as if they failed to generalize when they were in environments in which the incidental cue was absent. Subsequently, when the second teacher simply introduced the incidental cue (e.g., raised a hand in a

similar way) in the outside setting, the children did generalize appropriately. That is, generalized responding occurred only after the systematic exploration and isolation of the controlling stimuli in the treatment environment and the introduction of these specific stimuli to new (outside) environments.

Another study, designed to help understand the autistic child's problem with generalizing social stimuli, can also be understood as a problem caused by stimulus overselectivity. Specifically, Schreibman and Lovaas (1973) taught normal and autistic children to discriminate between lifelike male and female figures (dolls). Subsequent tests showed that normal children distinguished between the figures on the basis of a number of cues, including the figures' heads; but autistic children used only some minor and unreliable feature, such as the figures' shoes, and did not respond to reliable cues such as the figures' heads. For example, when the investigators removed the shoes from the figures after a child had learned to tell the figures apart, the child suddenly lost the discrimination, only to regain it in additional training by using an equally unreliable feature. Parents of autistic children often report that even a minor change in their (the parents') appearance, such as a mother cutting her hair or the father removing his glasses, may cause an abrupt change in their child's behavior, such as treating the parents as strangers or responding with major emotional upheaval. Perhaps the problem autistic children have with stimulus overselectivity contributes to their social aloofness in the presence of the complex, multidimensional stimuli provided by human beings.

Remedial Treatment

Within-Stimulus Prompts

The studies described above may help to explain why autistic children fail to develop adequately. The main strength of these studies, however, lies in the guidance they provide in construction of more therapeutic environments for such children. For example, the autistic child's problems with situations that involve prompts may lead to the design of therapeutic learning environments that prevent the deleterious effect of such prompts. The therapist or teacher may avoid situations in which the autistic child would overselect a nondistinctive (i.e., irrelevant) stimulus element and thereby fail to learn the discrimination. Two recent studies investigated the possibility of preventing the development of such unintended stimulus–response relationships. They employed procedures designed to establish control by the distinctive feature of the correct stimulus at the start of training and to maintain this control while irrelevant features were gradually faded in.

In the first study, Schreibman (1975) wanted to teach autistic children the difference between form stimuli that were difficult to tell apart. A child's everyday environment abounds with difficult form discriminations, such as the difference between people who smile and frown, a heater dial being turned on and off, letters like *b* and *d*, and so on. Normal development seems to require the acquisition of such discriminations. In one of the discriminations employed by Schreibman, the children were required to identify one of two very similar stick figures, shown in Figure 2. The figures were identical except that one had an arm raised, whereas the other had both arms down. Thus, the only relevant component of the discrimination was the orientation of the arms; all other components were redundant. The child was asked to point to the "correct" figure. For the purposes of this experiment, for some of the children the figure with the raised arm was designated correct; for others the other figure was correct. The autistic children failed to learn this discrimination even though the teacher went through extensive and elaborate prompting procedures, such as carefully pointing to the correct stimulus and then gradually removing the pointing cues. The autistic children did learn to discriminate between increasingly fine "finger points," but did not transfer their response from the teacher's finger to the stick figures. Schreibman then used a different prompt procedure that involved altering and exaggerating the relevant component of the discrimination (arm orientation) and prompting within the same stimulus dimension as the training stimulus. Figure 2 illustrates this procedure. The children were initially pre-

sented with two cards. One card was blank, and the other card had a heavy diagonal line across its surface. Step 1 of Figure 2 shows this discrimination. As the children learned this discrimination, size cues were gradually faded out (see Figure 2) until the children were responding to the line orientation (relevant cue). Then, the redundant components of the discrimination (head, body) were slowly faded in. Using this within-stimulus prompt fading procedure, the autistic children learned the discrimination that they had failed to learn with traditional prompting. The main difference in procedure was that the within-stimulus prompt did not require the children to respond to multiple cues, since the prompt was within the training stimulus.

Rincover (1978) extended Schreibman's work with an analysis of four visual fading procedures and their effectiveness in teaching autistic children discriminations between three-letter words. Two variables were assessed: (a) distinctive versus nondistinctive feature fading, which signified whether a prompt was a feature contained only in the S^D or contained in both the S^D and S^Δ, and (b) within- versus extrastimulus fading, which signified whether the prompt was superimposed on the S^D during fading or presented spatially separate from the S^D. The four fading procedures were all combinations of these two variables. As expected, the combination of distinctive-feature and within-stimulus prompting was most effective in teaching the discriminations. It was suggested that this success was due to the fact that response to the stimulus component first encountered in training was all that was required of the child; this component was never removed during the fading progression and was still available in the criterion discrimination. Thus, neither response to multiple cues nor a shift in stimulus control was necessary, as they were in the other three training procedures.

Multiple-Cue Training

Techniques such as those suggested by Schreibman (1975) and Rincover (1978) work around the overselectivity problem by constructing learning environments that allow the child to remain overselective, yet learn. However, it may also be possible to work directly on the overselectivity problem, assuming it is not permanent but itself amenable to learning. In our clinical work, as in the work of others (e.g., Risley & Wolf, 1967), autistic children eventually learned to "use" extrastimulus prompts rather than always be hindered by them. Although the exact reasons for this were not evident in the clinical work per se, the results of recent research suggest that overselectivity may be modifiable.

Schover and Newsom (1976) directly attacked the problem of overselectivity by training autistic children to broaden their

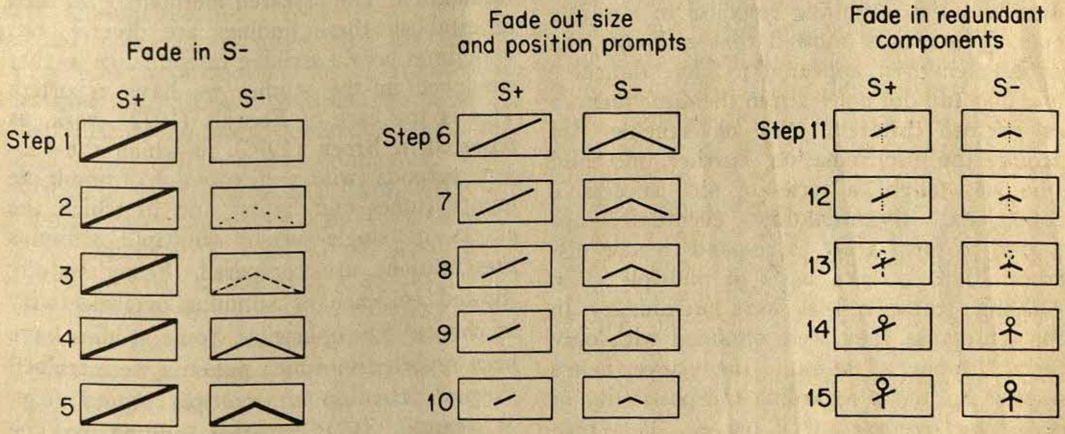

Figure 2. The within-stimulus prompt and fading steps used to teach discrimination between two stick figures. (Adapted from "Effects of Within-Stimulus and Extra-Stimulus Prompting on Discrimination Learning in Autistic Children" by L. Schreibman, *Journal of Applied Behavior Analysis*, 1975, *8*, 91–112. Copyright 1975 by the Society for the Experimental Analysis of Behavior, Incorporated. Reprinted by permission.)

responding to include multiple cues. They examined the effects of overtraining an already learned discrimination between simple figures (such as a large green square and a small orange triangle). Their data showed that overtraining increased the number of cues responded to by autistic children. They concluded that the overselectivity shown by autistic children is probably not the result of a permanent disability, or that it may be permanent but treatable, like diabetes.

A study by Schreibman, Koegel, and Craig (1977) further strengthens this inference. Using discrimination tasks, they found that simple overtraining (just exposure) did not help the autistic child to respond to more cues. However, prolonged testing with unreinforced probe trials interspersed among reinforced training trials eliminated overselective responding in 13 of 16 autistic children who were initially overselective. That is, children who selectively responded to one of the two available cues early in testing eventually responded to both cues as test trials continued. Since all of the stimuli in this study were in the visual modality, it is possible that the rapid reduction in overselectivity was a function of the stimuli used.

In perhaps the most direct attempt to test the hypothesis that overselectivity could be modified, Koegel and Schreibman (1977) taught five autistic children a conditional discrimination requiring response to multiple cues. The results showed that although the autistic children appeared to have difficulty learning and did not learn in the same manner as normal children, they nevertheless did acquire the discrimination. Further, one child who was taught a series of nine successive conditional discriminations eventually appeared to form a set to respond to new discriminations on the basis of multiple cues. Although these results were preliminary in the sense that they were obtained with only certain types of stimuli, they nevertheless suggest optimism regarding the possibility of correcting overselectivity per se. This possibility has numerous implications. Perhaps the most important are that (a) if overselectivity were eliminated, autistic children might be more likely to benefit from more traditional teaching procedures that typically require such responding, and (b) elimination of overselectivity may enable the children to respond to their environment in a more normal manner and thus facilitate large-scope, rapid changes in behavior.

These studies on eliminating the deleterious effects of stimulus overselectivity suggest that such overselectivity may be overcome or taken advantage of, as the situation requires, if sufficient effort and imagination are used in teaching autistic children. The findings provide basis for optimism that an experimentally derived technology for successfully teaching autistic children could be forthcoming.

Relation to Adult Schizophrenia, Learning Disabilities, and Mental Retardation

The discussion of overselective responding in autism shows a striking similarity to certain research on adult schizophrenics. Researchers in the area of adult schizophrenia typically distinguish between acute and chronic schizophrenics on the basis of breadth of cue utilization. Acute schizophrenics are described as responding to too many cues in their environment. They are not able to select out the relevant stimulation to the exclusion of the irrelevant. In contrast, chronic schizophrenics are more like autistic children in that they respond to a limited amount of available stimulation. The research methodologies used to establish these findings are diverse, but sometimes bear a striking resemblance to that employed in the studies we have reported. This is the case of Feeny's (1972) work, as reported in Broen (1973), in which the subject responds (with a microswitch) to multiple inputs (tones and lights) and in which the effects of single versus multiple stimulus presentations are compared. Feeny's study showed evidence of stimulus overselectivity in chronic schizophrenics. Some studies have been reported in which patients were trained to pay attention to multiple inputs (e.g., Meiselman, 1971), further adding to the similarity in our approaches. Broen's (1973) description of the findings on limited or narrowed attention in chronic schizophrenics seems relevant to the research we have

reviewed on autism:

> Chronic schizophrenics do indeed show a narrower range of cue utilization than normals or acute schizophrenics. Their ability to note and respond to relevant cues is especially impaired when the cues are located in more than one sensory modality. (p. 207)

In the same article, Broen speculates on the effects of such restricted responding:

> A life style dedicated to limiting stimulation . . . increases the likelihood of a *chronic* inability to monitor the environment and adjust to its changing demands . . . [it] maintains the potential for disorganization in the face of complexity. (p. 193)

If we consider the fact that many autistic children become diagnosed as chronic schizophrenics as adults, then our data may help to better understand the process of chronic schizophrenia. This seems particularly true if one assumes that children with a less complex history would show such problems as underlie schizophrenia in a more "pure" form. In any case, it seems striking as well as encouraging that two independent areas of investigation have produced such similar findings and reached such similar conclusions; one must, however, exercise considerable caution in drawing inferences common to adult chronic schizophrenia and autism, considering the problems in diagnosis and the diversity in research methodology. Other reviews of the literature on cue utilization and schizophrenia have been provided by Lang and Buss (1965), McGhie and Chapman (1961), Silverman (1964), and Venables (1964).

The concepts discussed here also appear similar to those discussed in the literature on learning disabilities in children. People in this field discuss auditory dominance (e.g., Senf & Treundl, 1971), visual dominance (Baker & Raskin, 1973; Gaines & Raskin, 1970), and sensory integration (e.g., Baker & Raskin, 1973; Birch & Belmont, 1964; Chalfant & Scheffelin, 1969). Indeed a great deal of research in this area points to defective attention as central to learning disabilities (e.g., Dykman, Ackerman, Clements, & Peters, 1971; Luria, 1961; Ross, 1976; Senf & Treundl, 1971; Strauss & Lehtinen, 1947; Trabasso & Bower, 1968; Zeaman & House, 1963). The possibility exists, however, that the results applicable to learning disabled children may not be directly comparable to the results with autistic children because of the many definitions of "learning disabled" and because of different methods of investigation. For example, Tarver et al. (1976) reported that some learning disabled children may be underselective and respond to too many cues. However, Ross (1976) has pointed out several similarities between the difficulties manifested by autistic children and certain learning disabled children when required to respond to cross-modality multiple cues. He reviews several studies supporting this interpretation (e.g., Vande Voort, Senf, & Benton, 1972).

In the area of mental retardation, investigators such as Zeaman and House (1963) have carried out extensive studies on various aspects of discrimination learning and have concluded that deficits in response to multiple cues may be a crucial factor in the poor performance of such children. When these investigators compared children of various levels of retardation, they found that the lower level retarded children took longest to learn. However, there were no differences between children in learning rates once improvement began. That is, rather than very slowly and gradually acquiring a discrimination, the more severely retarded children showed no increases in correct responding for many trials and then suddenly showed large increases in correct responding, at the same rate as normal children. The difference between the high- and low-level children was in how many trials they took before they started to show increases in correct responding. These investigators also found that they could shorten the number of trials it took the low-level children (i.e., decrease the differences between high- and low-level children) by directing their attention to a relevant cue in the task. The general conclusion drawn from experiments such as these has been that retarded children take a long time to learn because their attention is not focusing on a relevant aspect of the discrimination but that once they attend to a relevant cue, they can learn as fast as normal children. Relating this conclusion to our results, we would speculate that the reason retarded children take so long to attend to a relevant cue is that they are sampling (responding to) fewer cues at a time than normal children, and therefore the

probability is lower that a relevant cue will be included in any given sample.

Relation to Other Work on Attentional Problems in Autism

Attentional abnormalities have been suspected by many authors on autism as the basic problem in the etiology of that disorder. The attentional problem is well illustrated in Kanner's (1944) description of one of his patients:

> When spoken to, he went on with what he was doing as if nothing had been said. Yet one never had the feeling that he was willingly disobedient or contrary. He was obviously so remote that the remarks didn't reach him. (p. 212)

The clinical literature on autism abounds with similar descriptions, as when the parents suspect the child may be blind and deaf only to be puzzled at observing that sometimes their child can see and hear quite well. For example, the child may not respond to his or her name being called or startle at a loud sound like a slamming door, but may respond to a barely audible siren or the sound of candy being unwrapped. Koegel and Schreibman (1976) tell of a child who showed this inconsistency in responding to a remarkable degree. It may be this inconsistency or variability in responding that has led people to infer that the autistic child has a deficiency in perceptual or attentional mechanisms rather than a sensory deficit.

The following is a brief discussion of the major theories that have been advanced to account for the attentional deficits in autism. It is surprising how many of these theories postulate some restriction in attention, analogous to the stimulus overselectivity hypothesis that we have proposed, as basic to autistic development.

Psychodynamic Theories

Psychodynamic theorists presented the first attempt to account for the autistic child's unresponsivity, seeing such unresponsiveness as a defense against a hostile, threatening world. For example, Bettelheim (1967) proposed that autistic children narrow the range of their attention because they feel it necessary to shut out of awareness their disappointing, unresponsive, and destructive parents. It seems clear that the psychodynamic orientations talk of restricted or narrowed attention in the autistic child and that this is seen as a defense against interpersonal trauma.

Empirical research designed to investigate attentional deficits based on social trauma has provided little support for the psychodynamic formulations. For example, Hermelin and O'Connor (1963) studied autistic and retarded children (matched on IQ) in a free-field situation designed to assess the amount of attention to visual, auditory, manipulative, or social stimuli and found no significant difference between the groups in the amount of attention to the different stimuli. In a later study O'Connor and Hermelin (1963) found no evidence of autistic withdrawal from social stimuli. These same investigators (O'Connor & Hermelin, 1967) compared visual fixations of normal, severely retarded, and psychotic children to two simultaneous card displays depicting social (faces) and nonsocial stimuli. They found that (a) compared to the other groups, the psychotic children spent less time looking at any display cards and more time in nondirected gazing, and (b) all children, including psychotics, looked more at a picture of a face than at the scrambled pictures of the same face. Similar findings are reported by Young (1970). Thus, the experimental evidence does not lend support to the notion that autistic children display a particular attentional deficit in relation to social stimuli.

Developmental Theories

Developmental theories of perceptual deviations in autism are based on observations of animal and normal human development (Sherrington, 1906; Zaporozhets, 1961), and hold that there is a normal transition from preference and dependence on the near receptors (tactile, kinesthetic, gustatory) in early life to the dominance of input from the far receptors (visual, auditory) in later life. It has been postulated that the autistic child's unresponsiveness to auditory and visual stimulation is due to a failure in development beyond the near-receptor stage (Goldfarb, 1956; Schopler, 1965).

Some data are consistent with this model. For example, Pollack and Goldfarb's (1957) data suggested that autistic children were not making good use of added visual cues. Frith and Hermelin (1969) presented additional evidence consistent with these findings. They studied normal, autistic, and retarded children to compare the relative use of visual and tactile cues. For example, in one experiment the children were tested on three tasks. One task could be solved on the basis of visual cues only. The second task could be solved by either visual or tactile cues or both. The third task could only be solved using tactile cues. They found that all children did better with the visual cues except the developmentally backward autistic children. For them, providing visual information had no facilitatory effect. Goldfarb and Braunstein (1958) presented similar data, as in their report that normal children's speech was disturbed while listening to delayed auditory feedback and the speech of childhood schizophrenics was unaffected.

Schopler (1966) found that normal and retarded subjects showed a higher preference for visual stimulation than did the schizophrenic subjects. (Normals and schizophrenics were matched on chronological age [CA], retardates and schizophrenics on mental age [MA].) He also found that normal subjects increased in their visual preference with age. Similarly, Hermelin and O'Connor (1964) presented psychotic and normal children of the same MA and CA with stimuli in different modalities and found that the psychotic children responded much more often to touch and the normal children to sound.

But much empirical research has failed to support the developmental model of sensory deficit. In Schopler's (1966) study he did not find a difference between the normal, retarded, and schizophrenic subjects in their preference for tactile cues, nor did he find that the normals decreased in their preference for tactile stimulation with age. Also, the retarded subjects in that study showed an increase in tactile preference with a decrease in visual preference with increase in MA. Both these findings are difficult to fit into a developmental interpretation. Also, Goldfarb (1961) found no significant difference in the abilities of normal and schizophrenic children in relation to visual, auditory, and tactile cues. In the Hermelin and O'Connor study (1963) cited earlier, autistic and retarded children did not differ in their responsiveness to visual, auditory, and tactile stimuli. The same investigators (O'Connor & Hermelin, 1965) failed to confirm their earlier (Hermelin & O'Connor, 1964) findings of near-receptor preference in autistics. Finally, in our own studies (e.g., overselective response to three cues and to two cues) we failed to observe the autistic's preference for a particular stimulus hierarchy. The autistic subjects responded on the basis of either visual or auditory cues, and none of them responded to the tactile (near-receptor) cue.

In general, the literature on developmental receptor preferences does not provide a tenable explanation of sensory deficit in autism. But the data do support the notion that autistic children may act like a developmentally immature organism in their restricted response to multiple cues. Perhaps developmental variables may sometimes be understood as related to "breadth of cue use" independent of sensory preferences.

Arousal Theories

Several theories of sensory dysfunction in autism have emphasized the physiological mechanisms related to arousal. Since the level of physiological arousal is an important determinant of how much the organism will be affected by environmental stimulation, it is reasonable to postulate that pathological mechanisms influencing arousal would interfere with normal sensory processes. There are three basic arousal theories represented in the literature. One position suggests that autistic children suffer from a chronically low level of arousal, another suggests a chronically high level of arousal, and the third postulates alternating periods of high and low arousal.

The main proponent of the underarousal theory is Rimland (1964). He hypothesized that autistic children are chronically underaroused because of a dysfunction of the reticular activating system. A stage of underarousal would account for the child's restricted attention to external stimulation. Rimland's use (pp. 201–204) of "narrow bands" (from

information theory) seems particularly close to our use of stimulus overselectivity in accounting for the behavioral peculiarities of autism. Metz (1967) provided some data consistent with the underarousal hypothesis. He found that when autistic and normal subjects were allowed to control the volume of an auditory stimulus, the autistics preferred higher levels of stimulation than did the normals. But in general the underarousal theory has little direct supporting evidence and has not been subjected to rigorous empirical investigation.

The overarousal hypothesis is based on more empirical findings. Hutt, Hutt, Lee, and Ounsted (1965) hypothesized that in autistic children the nonspecific activity of the reticular activating system is sustained at a high and relatively inflexible level. They point to the possibility that the typical unresponsiveness displayed by autistics is a defensive function preventing excessive arousal. Hutt, Hutt, Lee, and Ounsted (1964) suggested that the overarousal hypothesis is also consistent with the desire for sameness in the environment. Novelty leads to increased arousal and is thus avoided. Some support for the overarousal position is provided by Connell (1966), who observed that autistics often require higher than normal doses of sedative drugs. Rutter (1968) described two important problems with an overarousal theory. First, level of arousal may be related to level of maturation or development. The Hutt et al. (1965) subjects were matched on CA but not MA. Thus it is possible that the lower MA of the autistics accounts for the differential electroencephalogram (EEG) patterns of the two groups. Second, high arousal may have developed as a secondary rather than a primary defect. Hermelin and O'Connor (1968) found no major differences in alpha rhythm in the EEGs of autistic, Down's syndrome, and normal children in several conditions differing in amount of sound and light stimulation. Only when continuous noise was present in one of the conditions did the autistics display relatively more arousal than the other groups. This suggests that overarousal may be a secondary response to environmental stimulation.

A third arousal theory is the "perceptual inconstancy" hypothesis offered by Ornitz and Ritvo (1968). According to this theory, there is a specific pathological process that begins in the first year of life when there is a failure of the central nervous system (CNS) to develop adequate homeostatic regulation of sensory input. Without control over incoming sensory stimulation, the child experiences random overloading and underloading of the CNS. At it is now formulated, the perceptual inconstancy hypothesis does not relate clearly to our work on stimulus overselectivity. Although both viewpoints address the problem of selective attention, our data suggest that the autistic children responded to only one stimulus component (auditory or visual) in a complex stimulus and that they responded to that same component consistently day after day. If, as Ornitz and Ritvo say, the underloading and overloading of the CNS is random, we should not have observed such consistent patterns of responding.

Other Relationships

Much may be learned by relating the studies on stimulus overselectivity more closely to similar work with animals. A thorough review of that work is beyond the scope of this article, but it should be mentioned that Pavlov noted in 1927 that the conditioned response to one element of a complex stimulus was as large as the response to the complex, leaving the response to the other elements negligible. Similar early references to the distinction between "nominal" and "functional" or effective stimuli can be found in the work of Harlow (1945) and Warren (1953). Reynolds (1961), for example, trained two pigeons to discriminate between two white forms on differently colored backgrounds (red or green). It was found that one pigeon responded only to the white form and the other pigeon responded only to the colored background. Some persons have considered the underlying mechanism behind selective responding to be genetic (some cues are more dominant for some species than others). Other persons have attributed selective responding to prior learning of one cue, which then "blocks" (as in "stimulus blocking") or inhibits responding to other cues that are also

available. Perhaps animal studies may provide information about the basic cause of (over) selective responding.

Conclusion

Looking across all of these studies, one can make several general observations. It is apparent that under a wide range of different testing situations, low-level autistic children come under the control of an extremely restricted range of stimuli. At this point, however, it may be premature to label this phenomenon in other than a descriptive manner. Thus we have chosen to employ the term *stimulus overselectivity*, rather than other more inferential terms such as *association deficiency* or *deficiency in selective attention*. There are, however, many parallels between our results and those discussed in the discrimination learning literature in the area of selective attention.

We are limited in our inferences regarding the mechanism behind the phenomenon, but we are still able to relate the children's overselective responding to a number of their abnormalities. Specifically, we have reviewed findings relating overselectivity to deficiencies in (a) generalization, (b) the use of "extra-stimulus" prompts, (c) language learning, (d) social behavior, and (e) observational learning. We have also been able to relate overselectivity to specific remedial procedures. For example, the work on prompting has resulted in an extremely efficient technique (within-stimulus prompting) for teaching at least some behaviors. Our major observation, and our main point in this article, is that overselectivity frequently occurs in children diagnosed as autistic, that this characteristic can be reliably measured, and that a knowledge of its existence may be useful in planning for these children's education and treatment.

Reference Note

1. Russo, D. C. *Variables influencing transfer from prompt to training stimuli in autistic children: Difficulty of the discrimination.* Unpublished manuscript, 1976. (Available from D. C. Russo, Behavioral Psychology, Children's Hospital Medical Center, 300 Longwood Avenue, Boston, Massachusetts 02115.)

References

Baker, G. F., & Raskin, L. M. Sensory integration in the learning-disabled. *Journal of Learning Disabilities*, 1973, *6*, 645–649.

Bandura, A. *Principles of behavior modification.* New York: Holt, Rinehart & Winston, 1969.

Bettelheim, B. *The empty fortress: Infantile autism and the birth of the self.* New York: Free Press, 1967.

Birch, H. G., & Belmont, L. Auditory-visual integration in normal and retarded readers. *American Journal of Orthopsychiatry*, 1964, *34*, 852–861.

Broen, W. E. Limiting the flood of stimulation: A protective deficit in chronic schizophrenics. In R. L. Solso (Ed.), *Contemporary issues in cognitive psychology: The Loyola Symposium.* Washington, D.C.: V. H. Winston, 1973.

Bryson, C. Q. Systematic identification of perceptual disabilities in autistic children. *Perceptual and Motor Skills*, 1970, *31*, 329–346.

Carr, E. G., Newsom, C. D., & Binkoff, J. A. Stimulus control of self-destructive behavior in a psychotic child. *Journal of Abnormal Child Psychology*, 1976, *4*, 139–153.

Chalfant, J. C., & Scheffelin, M. A. *Central processing dysfunction in children: A review of research* (NINDS Monograph No. 9). Washington, D.C.: U.S. Government Printing Office, 1969.

Connell, P. H. Medical treatment. In J. K. Wing (Ed.), *Childhood autism: Clinical, educational, and social aspects.* London: Pergamon Press, 1966.

Cowan, P. A., Hoddinott, B. A., & Wright, B. S. Compliance and resistance in the conditioning of autistic children: An exploratory study. *Child Development*, 1965, *36*, 913–923.

Dykman, R. A., Ackerman, P. T., Clements, S. D., & Peters, J. E. Specific learning disabilities: An attentional deficit syndrome. In H. R. Myklebust (Ed.), *Progress in learning disabilities* (Vol. 2). New York: Grune & Stratton, 1971.

Eimas, P. Multiple-cue discrimination learning in children. *Psychological Record*, 1969, *19*, 417–424.

Feeny, S. Breadth of cue utilization and ability to attend selectively in schizophrenics and normals (Doctoral dissertation, University of California, Los Angeles, 1971). *Dissertation Abstracts International*, 1972, *32*, 4208B. (University Microfilms No. 72–2810)

Fellows, B. J. *The discrimination process and development.* London: Pergamon Press, 1968.

Fischer, M. A., & Zeaman, D. An attention-retention theory of retardate discrimination learning. In N. R. Ellis (Ed.), *International review of research in mental retardation* (Vol. 6). New York: Academic Press, 1973.

Frith, U. M., & Hermelin, B. The role of visual and motor cues for normal, subnormal, and autistic children. *Journal of Child Psychology, Psychiatry, and Allied Disciplines*, 1969, *10*, 153–163.

Gaines, B. J., & Raskin, L. M. Comparison of cross-modal and intra-modal form recognition in children with learning disabilities. *Journal of Learning Disabilities*, 1970, *3*, 243–246.

Goldfarb, W. Receptor preferences in schizophrenic children. *Archives of Neurological Psychology*, 1956, *76*, 643–652.

Goldfarb, W. *Childhood schizophrenia*. Cambridge, Mass.: Harvard University Press, 1961.

Goldfarb, W., & Braunstein, P. Reactions to delayed auditory feedback among a group of schizophrenic children. In P. H. Hock & J. Zubin (Eds.), *Psychopathology of communication*. New York: Grune & Stratton, 1958.

Hale, G. A., & Morgan, J. S. Developmental trends in children's component selection. *Journal of Experimental Child Psychology*, 1973, *15*, 302–314.

Harlow, H. F. Studies in discrimination learning in monkeys: V. Initial performance by experimentally naive monkeys on stimulus and pattern discriminations. *Journal of General Psychology*, 1945, *33*, 3–10.

Hermelin, B., & O'Connor, N. The response and self-generated behavior of severely disturbed children and of subnormal controls. *British Journal of Social and Clinical Psychology*, 1963, *2*, 37–43.

Hermelin, B., & O'Connor, N. Effects of sensory input and sensory dominance on severely disturbed children and on subnormal controls. *British Journal of Psychology*, 1964, *55*, 201–206.

Hermelin, B., & O'Connor, N. Measures of the occipital alpha rhythm in normal, subnormal and autistic children. *British Journal of Psychiatry*, 1968, *114*, 603–610.

Hewett, F. M. Teaching speech to autistic children through operant conditioning. *American Journal of Orthopsychiatry*, 1965, *35*, 927–936.

Hintgten, J. N., & Churchill, D. W. Differential effects of behavior modification in four mute autistic boys. In D. W. Churchill, G. D. Alpern, & M. K. De Myer (Eds.), *Infantile autism: Proceedings of the Indiana University Colloquium*. Springfield, Ill.: Charles C Thomas, 1971.

Hutt, S. J., Hutt, C., Lee, D., & Ounsted, C. Arousal and childhood autism. *Nature*, 1964, *204*, 908–909.

Hutt, S. J., Hutt, C., Lee, D., & Ounsted, C. A behavioral and encephalographic study of autistic children. *Journal of Psychiatric Research*, 1965, *3*, 181–197.

Kanner, L. Autistic disturbances of affective contact. *Nervous Child*, 1943, *2*, 217–250.

Kanner, L. Early infantile autism. *Journal of Pediatrics*, 1944, *25*, 211–217.

Koegel, R. L., Firestone, P. B., Kramme, K. W., & Dunlap, G. Increasing spontaneous play by suppressing self-stimulation in autistic children. *Journal of Applied Behavior Analysis*, 1974, *7*, 521–528.

Koegel, R. L., & Rincover, A. Some detrimental effects of using extra stimuli to guide responding in autistic and normal children. *Journal of Abnormal Child Psychology*, 1976, *4*, 59–71.

Koegel, R. L., & Schreibman, L. Identification of consistent responding to auditory stimuli by a functionally "deaf" autistic child. *Journal of Autism and Childhood Schizophrenia*, 1976, *6*, 147–156.

Koegel, R. L., & Schreibman, L. Teaching autistic children to respond to simultaneous multiple cues. *Journal of Experimental Child Psychology*, 1977, *2* 299–311.

Koegel, R. L., & Wilhelm, H. Selective respondi to the components of multiple visual cues by autist children. *Journal of Experimental Child Psycholog* 1973, *15*, 442–453.

Lang, P. J., & Buss, A. H. Psychological deficit i schizophrenia: II. Interference and activation. *Jou nal of Abnormal Psychology*, 1965, *70*, 77–106.

Lovaas, O. I. Program for establishment of spee in schizophrenic and autistic children. In J. K. Wi (Ed.), *Early childhood autism: Clinical, education and social aspects*. London: Pergamon Press, 1966.

Lovaas, O. I. *The autistic child: Language developme through behavior modification*. New York: Irvingto 1977.

Lovaas, O. I., Freitas, L., Nelson, K., & Whelan, C The establishment of imitation and its use for th development of complex behavior in schizophreni children. *Behaviour Research and Therapy*, 1967, *5* 171–181.

Lovaas, O. I., Koegel, R. L., Simmons, J. Q., & Long J. S. Some generalization and follow-up measure on autistic children in behavior therapy. *Journal o Applied Behavior Analysis*, 1973, *6*, 131–165.

Lovaas, O. I., & Newsom, C. D. Behavior modification with psychotic children. In H. Leitenberg (Ed.) *Handbook of behavior modification and behavio therapy*. New York: Appleton-Century-Crofts, 1976.

Lovaas, O. I., & Schreibman, L. Stimulus overselectivity of autistic children in a two stimulus situation. *Behaviour Research and Therapy*, 1971, *9*, 305–310.

Lovaas, O. I., Schreibman, L., Koegel, R. L., & Rehm, R. Selective responding by autistic children to multiple sensory input. *Journal of Abnormal Psychology*, 1971, *77*, 211–222.

Luria, A. R. *The role of speech in the regulation of normal and abnormal behavior*. New York: Liveright, 1961.

Maltzman, I., & Raskin, D. C. Effects of individual differences in the orienting reflex on conditioning and complex processes. *Journal of Research in Personality*, 1965, *1*, 1–16.

McGhie, A., & Chapman, J. S. Disorders of attention and perception in early schizophrenia. *British Journal of Medical Psychology*, 1961, *34*, 103–114.

Meiselman, K. C. Training chronic nonparanoid schizophrenics in dual modality attention (Doctoral dissertation, University of California, Los Angeles, 1971). *Dissertation Abstracts International*, 1971, *32*, 6654B. (University Microfilms No. 72-13, 633)

Metz, J. R. Conditioning generalized imitation in autistic children. *Journal of Experimental Child Psychology*, 1965, *2*, 389–399.

Metz, J. R. Stimulation level preferences of autistic children. *Journal of Abnormal Psychology*, 1967, *72*, 529–535.

O'Connor, N., & Hermelin, B. Measures of distance and motility in psychotic children and severely subnormal controls. *British Journal of Social and Clinical Psychology*, 1963, *3*, 29–33.

O'Connor, N., & Hermelin, B. Sensory dominance in autistic imbecile children and controls. *Archives of General Psychiatry*, 1965, *12*, 99–103.

O'Connor, N., & Hermelin, B. The selective visual attention of psychotic children. *Journal of Child Psychology and Psychiatry*, 1967, *8*, 167–179.

Olson, D. R. Information-processing limitations of mentally retarded children. *American Journal of Mental Deficiency*, 1971, *75*, 478–486.

Ornitz, E. M., & Ritvo, E. R. Perceptual inconstancy in early infantile autism. *Archives of General Psychiatry*, 1968, *18*, 76–98.

Pavlov, I. P. *Conditioned reflexes* (G. V. Anrep, trans.). London: Oxford University Press, 1927.

Pollack, M., & Goldfarb, W. The face-hand test in schizophrenic children. *Archives of Neurological Psychiatry*, 1957, *77*, 635–642.

Ray, B. A. Strategy in studies of attention: A commentary on D. I. Mostofsky's "Attention: Contemporary theory and analysis." *Journal of the Experimental Analysis of Behavior*, 1972, *12*, 293–297.

Ray, B. A., & Sidman, M. Reinforcement schedules and stimulus control. In W. N. Schoenfeld (Ed.), *The theory of reinforcement schedules*. New York: Appleton-Century-Crofts, 1970.

Reynolds, B. S., Newsom, C. D., & Lovaas, O. I. Auditory overselectivity in autistic children. *Journal of Abnormal Child Psychology*, 1974, *2*, 253–263.

Reynolds, G. S. Attention in the pigeon. *Journal of the Experimental Analysis of Behavior*, 1961, *4*, 203–208.

Rimland, B. *Infantile autism*. New York: Appleton-Century-Crofts, 1964.

Rincover, A. Variables affecting stimulus fading and discriminative responding in psychotic children. *Journal of Abnormal Psychology*, 1978, *87*, 541–553.

Rincover, A., & Koegel, R. L. Setting generality and stimulus control in autistic children. *Journal of Applied Behavior Analysis*, 1975, *8*, 235–246.

Risley, T. R., & Wolf, M. M. Establishing functional speech in echolalic children. *Behaviour Research and Therapy*, 1967, *5*, 73–88.

Ross, A. O. *Psychological aspects of learning disabilities and reading disorders*. New York: McGraw-Hill, 1976.

Rutter, M. Concepts of autism: A review of research. *Journal of Child Psychology and Psychiatry*, 1968, *9*, 1–25.

Schopler, E. Early infantile autism and the receptor processes. *Archives of General Psychiatry*, 1965, *13*, 327–335.

Schopler, E. Visual versus tactile receptor preference in normal and schizophrenic children. *Journal of Abnormal Psychology*, 1966, *71*, 108–114.

Schover, L. R., & Newsom, C. D. Overselectivity, developmental level, and overtraining in autistic and normal children. *Journal of Abnormal Child Psychology*, 1976, *4*, 289–298.

Schreibman, L. Effects of within-stimulus and extra-stimulus prompting on discrimination learning in autistic children. *Journal of Applied Behavior Analysis*, 1975, *8*, 91–112.

Schreibman, L., & Koegel, R. L. Autism: A defeatable horror. *Psychology Today*, 1975, *8*, 61–67.

Schreibman, L., Koegel, R. L., & Craig, M. S. Reducing stimulus overselectivity in autistic children. *Journal of Abnormal Child Psychology*, 1977, *5*, 425–436.

Schreibman, L., & Lovaas, O. I. Overselective response to social stimuli by autistic children. *Journal of Abnormal Child Psychology*, 1973, *1*, 152–168.

Senf, G. M., & Treundl, P. C. Memory and attention factors in specific learning disabilities. *Journal of Learning Disabilities*, 1971, *4*, 94–106.

Sherrington, C. S. *The integrative action of the nervous system*. London: Cambridge University Press, 1906.

Sidman, M., & Stoddard, L. T. Programming perception and learning for retarded children. In N. R. Ellis (Ed.), *International review of research in mental retardation* (Vol. 2). New York: Academic Press, 1966.

Silverman, J. The problem of attention in research and theory in schizophrenia. *Psychological Review*, 1964, *71*, 352–379.

Sivertsen, B. Overselectivity, mental age and behavioral development: A comparison of normal and autistic children (Doctoral dissertation, University of California, Los Angeles, 1976). *Dissertation Abstracts International*, 1976, *37*, 2562B. (University Microfilm No. 76–25, 242)

Strauss, A. A., & Lehtinen, L. E. *Psychopathology and education of the brain injured child*. New York: Grune & Stratton, 1947.

Sutherland, N. S., & MacKintosh, N. J. *Mechanisms of animal discrimination learning*. New York: Academic Press, 1971.

Tarver, S. G., Hallahan, D. P., Kauffman, J. M., & Ball, D. W. Verbal rehearsal and selective attention in children with learning disabilities: A developmental lag. *Journal of Experimental Child Psychology*, 1976, *22*, 375–385.

Terrace, H. Stimulus control. In W. K. Honig (Ed.), *Operant behavior: Areas of research and application*. New York: Appleton-Century-Crofts, 1966.

Touchette, P. E. The effects of graduated stimulus change on the acquisition of a simple discrimination in severely retarded boys. *Journal of the Experimental Analysis of Behavior*, 1968, *11*, 39–48.

Touchette, P. E. Transfer of stimulus control: Measuring the moment of transfer. *Journal of the Experimental Analysis of Behavior*, 1971, *15*, 347–364.

Trabasso, T., & Bower, G. H. *Attention in learning: Theory and research*. New York: Wiley, 1968.

Vande Voort, L., Senf, G. M., & Benton, A. L. Development of audio-visual integration in normal and retarded readers. *Child Development*, 1972, *43*, 1260–1272.

Varni, J. W., Lovaas, O. I., Koegel, R. L., & Everett, N. L. An analysis of observational learning in autistic and normal children. *Journal of Abnormal Child Psychology*, 1979, *7*, 31–43.

Venables, P. H. Input dysfunction in schizophrenia. In B. A. Maher (Ed.), *Progress in experimental personality research*, 1964, *1*, 1-47.

Warren, J. M. Additivity of cues in visual pattern discrimination by monkeys. *Journal of Comparative and Physiological Psychology*, 1953, *46*, 484-488.

Wilhelm, H., & Lovaas, O. I. Stimulus overselectivity: A common feature in autism and mental retardation. *American Journal of Mental Deficiency*, 1976, *81*, 227-241.

Young, S. Visual attention in autistic and normal children: Effects of stimulus novelty, human attributes, and complexity (Doctoral dissertation, University of California, Los Angeles, 1969). *Dissertation Abstracts International*, 1970, 31, 922B. (University Microfilms No. 70-14, 340)

Zaporozhets, A. V. The origin and development of the conscious control of movements in man. In N. O'Connor (Ed.), *Recent soviet psychology*. London: Pergamon Press, 1961.

Zeaman, D., & House, B. J. The role of attention in retardate discrimination learning. In N. R. Ellis (Ed.), *Handbook of mental deficiency*. New York: McGraw-Hill, 1963.

Received June 8, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1255–1263

Two-Sample T^2 Procedure and the Assumption of Homogeneous Covariance Matrices

A. Ralph Hakstian, J. Christian Roed, and John C. Lind
University of British Columbia, Vancouver, Canada

Results of an empirical investigation of the robustness of Hotelling's two-sample T^2 test with respect to violation of the assumption of homogeneity of covariance matrices are presented. Empirical sampling distributions of the T^2 statistic were obtained from a large number of sets, each consisting of 2,000 samples drawn from multivariate normal parent populations. Average sample size (n), extent of inequality of sample sizes, number of variables (p), and degree of inequality of covariance matrices were combined into 108 different conditions. Actual proportions of values that exceeded nominal α levels are presented. For equal ns, the procedure is shown to be generally robust. With unequal ns, the procedure is shown to become increasingly less robust as covariance matrix heterogeneity and p increase. The results are related to earlier findings, and implications for the proper use of the T^2 procedure are noted.

Hotelling's (1931) two-sample T^2 procedure for multiple dependent variables is widely used in the behavioral sciences today, being the uniformly most powerful test in the case of two-group p-variate simultaneous comparison (Anderson, 1958, pp. 115–118). Textbooks dealing with multivariate statistical methods in the behavioral sciences (e.g., Harris, 1975; Morrison, 1976; Tatsuoka, 1971), however, typically provide the reader with at most a brief treatment of the robustness of the method. The one investigation that is often referred to is that of Ito and Schull (1964), who investigated the large-sample properties of the distribution of T^2, showing analytically that for equal and large sample sizes, heterogeneity of covariance matrices has no substantial effect on the probability of Type I error.

Although infrequently referred to, several studies on this topic have been conducted. Two articles other than Ito and Schull's (1964) contain analytical results: Mardia (1971), who examined the effects of multivariate nonnormality (in the face of which the multivariate analysis of variance [MANOVA] class of tests is relatively robust), and Pillai and Sudjana (1975), who examined the effects of covariance matrix heterogeneity on four test criteria. Although limited to the bivariate case, small ns, and unclear degrees of heterogeneity, their results showed modest departures from nominal α values for minor degrees of heterogeneity and more pronounced departures with greater heterogeneity.

Several empirical Monte Carlo studies have dealt with the robustness of the T^2 test. Chase and Bulgren (1971) examined nonnormality, an issue not dealt with in the present article. Hopkins and Clay (1963) examined heterogeneity of covariance matrices, although only for the bivariate case. Olson (1974), who compared six MANOVA criteria, dealt with only the equal-n case. These studies provided some insights into the robustness of the T^2 test, although none illuminated the issue fully.

One reasonably comprehensive empirical study was that by Holloway and Dunn (1967), who, in general, found that with equal ns and a fairly large ratio of subjects to dependent variables, the T^2 test is robust. As expected,

The research reported in this article was supported in part by Grant A9088 from the National Research Council of Canada.

The authors gratefully acknowledge the advice of Ingram Olkin during the execution of the study.

Requests for reprints should be sent to A. Ralph Hakstian, Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5.

with unequal *ns*, the test was found to be less robust being conservative, when the larger sample was drawn from the population with generally greater dispersions, and liberal in the opposite situation. The main defect in this study, however, at least for the present purposes, was the unrealistically different heterogeneity conditions (e.g., variables in Population 2 that had variances 10 or 100 times those in Population 1) and unequal-n conditions that were not disparate enough ($n_1:n_2$ ranging from 15:35 to 35:15). In addition, for $n_1 \neq n_2$, $n_1 + n_2$ was fixed at 50, thus preventing the systematic examination of overall sample size as a factor in the issue of robustness.

Although something is known therefore about the robustness of the T^2 test, it is not so clear at what degree of departure from optimal conditions (equality of sample *ns*, homogeneity of covariance matrices, etc.) the procedure manifests an unacceptable actual α level. The purposes of the present study were (a) to examine simultaneously all independent variables relevant to robustness so that a comprehensive picture of each of these variables could be obtained, in terms of both their main effects and their possible interactive effects; (b) to represent the independent variables so that they represented real-world behavioral data in terms of levels that are identifiable with characteristics of observable data variables; and (c) to distill the results necessary to provide guidelines for the proper use of the T^2 procedure in psychological research.

The Monte Carlo Investigation

Independent Variables

Sample size. To assure comparability among data sets with different numbers of variates, the ratio of average sample size, $(n_1 + n_2)/2$, to number of variates was manipulated. Two ratios were used: 3 subjects per variable and 10 subjects per variable.

Inequality of sample sizes. The ratios of the sizes of the two samples were also varied. Three such ratios were used: 1:1, 2:1, and 5:1.

Number of variables. Three levels of this factor were used: 2, 6, and 10.

Heterogeneity of covariance matrices. This factor was central in the present study. Heterogeneity of population covariance matrices can, of course, arise from many causes and in a vast number of forms and degrees. It seems true, however, that the most obvious occurrence of such heterogeneity is that in which a sample that is drawn from a population with normal dispersions of the variables is compared with one that is either implicitly or explicitly selected or restricted in some way.

In the present study, the authors attempted to generate data that would possess the same heterogeneity often found in real-world data by introducing scale factors in four ways:

1. The variables in Population 2 were rescaled by 1.2, that is, whereas the variables in Population 1 were all set with $\sigma = 1$, those in Population 2 had σs of 1.2 and thus σ^2s of 1.44. This can be seen as a mild departure from homogeneity and is referred to as Heterogeneity Condition 1.
2. The variables in Population 2 were rescaled by 1.5. This scaling (Heterogeneity Condition 2) represented a moderate to substantial degree of heterogeneity. The Population 2 variances and covariances were thus 2.25 times those for Population 1.
3. The first $p/2$ variables in Population 2 were of the same scale as those in Population 1, whereas the last $p/2$ variables were rescaled by 1.5. Heterogeneity Condition 3 represented the situation of selection on some but not all of the variables.
4. The variables in the two populations were in precisely the same scale, that is, no rescaling took place. This is, of course, the homogeneity condition.

Population 1 covariance matrices (Σ_1) were constructed for the three p levels, 2, 6, and 10. These Σ_1 matrices appear in Table 1. Population 2 covariance matrices for each value of p (Σ_2) were then obtained by $\Sigma_2 = \mathbf{D}_i\Sigma_1\mathbf{D}_i$, where $i = 1, \ldots, 4$, and where the diagonal $\mathbf{D}_i$ matrices contained the scale factors referred to above. Thus, for Heterogeneity Conditions 1 and 2, $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively, contained 1.2 and 1.5 in each diagonal position. For Heterogeneity Condition 3, $\mathbf{D}_3$ contained 1 in the first $p/2$ diagonal positions and 1.5 in the last $p/2$, and for the homogeneity condition, $\mathbf{D}_4$ was simply $\mathbf{I}_p$.

Table 1
Population Covariance Matrices Used in the Study as Σ_1

2 variables: Variable	1	2
1	100	
2	50	100

6 variables: Variable	1	2	3	4	5	6
1	100					
2	40	100				
3	30	35	100			
4	40	25	10	100		
5	30	40	50	30	100	
6	20	30	00	40	15	100

10 variables: Variable	1	2	3	4	5	6	7	8	9	10
1	100									
2	40	100								
3	30	50	100							
4	50	30	30	100						
5	40	40	50	20	100					
6	20	20	40	50	40	100				
7	10	40	00	30	30	50	100			
8	20	30	50	10	50	30	40	100		
9	30	20	30	20	30	20	50	30	100	
10	50	00	10	40	10	40	30	00	20	100

Note. Entries are actual values multiplied by 100. The Σ_2 matrices that were used in the study were obtained from those in Table 1 by means of rescalings described in the text.

Previous treatments, such as those of Ito and Schull (1964) and Pillai and Sudjana (1975), have operationalized covariance matrix heterogeneity in terms of the diagonal matrix of latent roots yielded by the product $\Sigma_2\Sigma_1^{-1}$. Using the scheme for generating covariance matrix heterogeneity noted previously, it can easily be shown that since $\Sigma_2 = \mathbf{D}_i\Sigma_1\mathbf{D}_i$, the eigenstructure of $\Sigma_2\Sigma_1^{-1}$ or, equivalently, of $\mathbf{D}_i\Sigma_1\mathbf{D}_i\Sigma_1^{-1}$ contains a diagonal matrix of latent roots equal simply to $\mathbf{D}_i^2$. In Heterogeneity Conditions 1 and 2, these roots will, of course, all be equal.

Relationship between heterogeneity and inequality of sample size. This concerns whether, when $n_1 \neq n_2$, the larger sample was drawn from the population with larger dispersions (the positive condition) or smaller dispersions (the negative condition). Thus this two-level factor was not fully crossed with all other factors: It did not apply when $n_1 = n_2$ or when $\Sigma_1 = \Sigma_2$.

Summary of the independent variables studied. From the preceding, it can be seen that in all, 108 conditions were examined—90 conditions in which $\Sigma_1 \neq \Sigma_2$ [$2 \times 3 \times 5$ (2 positive $n_1{:}n_2$, 2 negative $n_1{:}n_2$, and 1 equal) $\times$ 3] and 18 in which $\Sigma_1 = \Sigma_2$ ($2 \times 3 \times 3$). An understanding of the overall design of the study will be facilitated by examination of Tables 2, 3, and 4, which appear later in the article.

Data Generation

The generation of sample data was accomplished by random number generation on the University of British Columbia IBM 370/168 computer. Independent uniformly distributed random numbers on the interval (0, 1) were generated and then transformed to normally distributed random numbers with mean 0 and variance 1 by Marsaglia's rectangular-wedge-tail method (Knuth, 1968). Strings of length Np (where $N = n_1 + n_2$) of such independent random normally distributed data points were generated and then partitioned into two data sets X (one, n_1 subjects by p variates, the other n_2 subjects by p variates). With the string partitioned in this way, each $n_j \times 1$ ($j = 1, 2$) variate vector was normally distributed, with mean 0 and

Table 2
Proportion of Values (Actual α Values) of the Two-Sample T^2 Statistic Exceeding the Nominal α Values .01, .05, and .10 Under Various Conditions for the Case of p = 2 Dependent Variables and a True Null Hypothesis

			Σ heterogeneity					
			Heterogeneity 1[a]		Heterogeneity 2[a]		Heterogeneity 3[a]	
$n_1:n_2$[b]	Nominal α	Homogeneity	Positive[c]	Negative[c]	Positive[c]	Negative[c]	Positive[c]	Negative[c]
			Average n per sample = 6					
6:6	.01	.010	.011		.012		.012	
	.05	.049	.052		.050		.054	
	.10	.104	.098		.102		.101	
8:4	.01	.008	.007	.021	.006	.023	.008	.017
	.05	.047	.030	.073	.031	.104	.035	.077
	.10	.098	.077	.133	.062	.175	.081	.142
10:2	.01	.009	.007	.020	.003	.045	.007	.035
	.05	.047	.027	.084	.014	.148	.029	.111
	.10	.101	.057	.147	.033	.256	.070	.181
			Average n per sample = 20					
20:20	.01	.011	.007		.013		.011	
	.05	.055	.050		.065		.046	
	.10	.105	.102		.113		.094	
27:13	.01	.010	.008	.019	.004	.029	.006	.016
	.05	.049	.045	.073	.026	.087	.034	.073
	.10	.097	.080	.134	.054	.155	.075	.127
33:7	.01	.016	.003	.025	.001	.058	.005	.035
	.05	.060	.025	.097	.007	.163	.031	.107
	.10	.111	.050	.170	.028	.238	.061	.185

Note. Each proportion is based on 2,000 pairs of samples. The various conditions involve Σ_1 and Σ_2 and magnitude and departure from equality of n_1 and n_2.
[a] The degree of heterogeneity present is discussed in the text.
[b] The $n_1:n_2$ ratios are such that they represent, as closely as possible, three levels: 1:1, 2:1, and 5:1.
[c] The positive condition refers to that in which the population whose covariance matrix contains the larger entries is that from which the larger n is drawn, where $n_1 \neq n_2$. The negative condition is that in which the population whose covariance matrix contains the larger entries is that from which the smaller n is drawn. For situations where $n_1 = n_2$, this dichotomy does not, of course, exist, and entries in Table 2 are given between the Positive and Negative columns.

variance 1 and was independent of every other vector. It can easily be demonstrated that the joint distribution thus arising is MVN(0, **I**) (see, for example, Anderson, 1958, pp. 19–27).

So that each data matrix **Y** would represent a sample from a population with a known covariance matrix Σ, the following transformation was applied. The desired population covariance matrix was first canonically decomposed as $\Sigma = \mathbf{V}\Lambda^2\mathbf{V}'$, and a "factor" matrix **F** was obtained by $\mathbf{F} = \mathbf{V}\Lambda$. Next, the $n_j \times p$ MVN(0, **I**) data matrices **X**, described previously, were postmultiplied by $\mathbf{F}'$, yielding $\mathbf{Y} = \mathbf{XF}'$. This process can be considered to produce a sample data matrix that could have arisen from a population having covariance matrix Σ, since

$$E(\mathbf{Y}'\mathbf{Y}) = E[(\mathbf{XF}')'(\mathbf{XF}')] = E(\mathbf{FX}'\mathbf{XF}')$$
$$= \mathbf{F}E(\mathbf{X}'\mathbf{X})\mathbf{F}' = \mathbf{FF}' = \mathbf{V}\Lambda^2\mathbf{V}' = \Sigma.$$

In this way, 2,000 pairs of samples (from populations in which the null hypothesis was

true) were generated for each of the 108 conditions studied. For each pair of samples, the T^2 statistic was computed and transformed to a value (F), which under a true null hypothesis is distributed as a central F variate with degrees of freedom p and $(n_1 + n_2 - p - 1)$. The obtained F values were then compared to the 90th, 95th, and 99th percentile points of the appropriate F distribution, and the percentage lying above those points was tabulated for each condition.

Results

The results of the described Monte Carlo analyses appear in Tables 2, 3, and 4; each table deals with a separate p value (2, 6, or 10). The tabled proportions, based as they all are on 2,000 sample pairs, are subject to mild sampling error and should be interpreted accordingly. Using the standard error of a proportion and normal curve probabilities, we can set approximate .95 confidence intervals

Table 3

Proportion of Values (Actual α Values) of the Two-Sample T^2 *Statistic Exceeding the Nominal α Values .01, .05, and .10 Under Various Conditions for the Case of* p = 6 *Dependent Variables and a True Null Hypothesis*

			Σ heterogeneity					
			Heterogeneity 1[a]		Heterogeneity 2[a]		Heterogeneity 3[a]	
$n_1:n_2$[b]	Nominal α	Homogeneity	Positive[c]	Negative[c]	Positive[c]	Negative[c]	Positive[c]	Negative[c]
			Average n per sample = 18					
18:18	.01	.013	.006		.011		.012	
	.05	.048	.048		.057		.064	
	.10	.098	.099		.109		.114	
24:12	.01	.011	.007	.020	.005	.043	.006	.018
	.05	.059	.035	.088	.021	.127	.028	.076
	.10	.106	.068	.155	.051	.214	.072	.158
30:6	.01	.011	.004	.036	.000	.103	.003	.046
	.05	.057	.018	.117	.004	.249	.022	.145
	.10	.097	.045	.202	.012	.358	.046	.231
			Average n per sample = 60					
60:60	.01	.009	.012		.016		.008	
	.05	.047	.056		.059		.040	
	.10	.096	.097		.111		.078	
80:40	.01	.011	.003	.015	.002	.040	.007	.021
	.05	.050	.029	.074	.011	.137	.026	.079
	.10	.098	.065	.141	.030	.225	.059	.141
100:20	.01	.010	.003	.038	.000	.132	.003	.067
	.05	.044	.015	.123	.003	.285	.021	.179
	.10	.096	.035	.202	.008	.393	.046	.258

Note. Each proportion is based on 2,000 pairs of samples. The various conditions involve Σ_1 and Σ_2 and magnitude and departure from equality of n_1 and n_2.

[a] The degree of heterogeneity present is discussed in the text.

[b] The $n_1:n_2$ ratios are such that they represent three levels: 1:1, 2:1, and 5:1.

[c] The positive condition refers to that in which the population whose covariance matrix contains the larger entries is that from which the larger n is drawn, where $n_1 \neq n_2$. The negative condition is that in which the population whose covariance matrix contains the larger entries is that from which the smaller n is drawn. For situations where $n_1 = n_2$, this dichotomy does not, of course, exist, and entries in Table 3 are given between the Positive and Negative columns.

Table 4
Proportion of Values (Actual α Values) of the Two-Sample T^2 *Statistic Exceeding the Nominal α Values .01, .05, and .10 Under Various Conditions for the Case of* p = *10 Dependent Variables and a True Null Hypothesis*

			Σ heterogeneity					
			Heterogeneity 1[a]		Heterogeneity 2[a]		Heterogeneity 3[a]	
$n_2:n_2$[b]	Nominal α	Homogeneity	Positive[c]	Negative[c]	Positive[c]	Negative[c]	Positive[c]	Negative[c]
			Average n per sample = 30					
30:30	.01	.009	.009		.010		.013	
	.05	.059	.049		.056		.058	
	.10	.104	.106		.109		.101	
40:20	.01	.008	.004	.020	.000	.058	.004	.026
	.05	.057	.029	.088	.013	.158	.023	.106
	.10	.100	.068	.158	.035	.251	.057	.183
50:10	.01	.009	.002	.041	.000	.152	.004	.081
	.05	.050	.015	.124	.005	.337	.027	.203
	.10	.097	.028	.235	.010	.473	.057	.303
			Average n per sample = 100					
100:100	.01	.011	.011		.013		.012	
	.05	.048	.049		.050		.050	
	.10	.112	.098		.099		.104	
133:67	.01	.007	.003	.020	.001	.054	.006	.026
	.05	.049	.022	.085	.010	.163	.036	.100
	.10	.100	.055	.157	.024	.256	.061	.175
167:33	.01	.014	.001	.050	.000	.187	.007	.084
	.05	.057	.008	.156	.001	.369	.025	.211
	.10	.107	.026	.241	.003	.484	.055	.311

Note. Each proportion is based on 2,000 pairs of samples. The various conditions involve Σ_1 and Σ_2 and magnitude and departure from equality of n_1 and n_2.
[a] The degree of heterogeneity present is discussed in the text.
[b] The $n_1:n_2$ ratios are such that they represent, as closely as possible, three levels: 1:1, 2:1, and 5:1.
[c] The positive condition refers to that in which the population whose covariance matrix contains the larger entries is that from which the larger n is drawn, where $n_1 \neq n_2$. The negative condition is that in which the population whose covariance matrix contains the larger entries is that from which the smaller n is drawn. For situations where $n_1 = n_2$, this dichotomy does not, of course, exist, and entries in Table 4 are given between the Positive and Negative columns.

around each nominal α value tabled as (a) .10 ± .013, (b) .05 ± .010, and (c) .01 ± .004. Thus, any value between, for example. .040 and .060 can be considered to be within sampling error of the nominal value of .050.

The results for the two-variable case appear in Table 2. We see, first, that the robustness of the T^2 test with equal ns extends to relatively small samples. As expected, the test is unaffected by inequality of ns with equal Σs. When sample sizes are unequal, however, and $\Sigma_1 \neq \Sigma_2$, the actual α level can differ considerably from nominal, falling below nominal in the positive case ($n_2 > n_1$) and exceeding it in the negative ($n_1 > n_2$). With the $n_1:n_2$ ratio fixed, the degree of departure from the nominal α level increases as the degree of inequality of covariance matrices increases. Also, with the degree of covariance matrix heterogeneity fixed, the degree of discrepancy between actual and nominal α values increases as the $n_1:n_2$ ratio departs from one. And in-

creasing the sample size while maintaining the ratio between the two sample sizes does not help; if anything, the reverse is true.

Tables 3 and 4 contain the results for, respectively, the 6- and 10-variable conditions. All of the earlier results generalize to conditions involving more variables. Again for a fixed $n_1:n_2$ ratio and degree of Σ heterogeneity, actual α values tended to be more discrepant for the larger average sample sizes than for the smaller.

Examination of Tables 2, 3, and 4 reveals that with an increase in the number of dependent variables, p, the effects of these factors on the statistical test become more pronounced, in spite of the fact that the ratio of average sample size to number of variates was held constant (3:1 and 10:1). In the most extreme cases reported here, the departure from nominal α levels became large indeed, as is seen, for example, in Table 4.

Conclusions

From these results, it is clear that the T^2 procedure is generally robust with respect to violation of the homogeneity of covariance matrix assumption for equal sample sizes, even when the ratio of sample size to number of dependent variables is small, for example, three subjects per variable. Under these conditions, the T^2 procedure has been shown to be able to withstand population scale differences of 1.5 on all variables (Heterogeneity Condition 2)—with the number of such variables as large as 10—a scale factor that implies a between-populations variance difference of 2.25, a difference that seems like a realistic extreme for behavioral data. Holloway and Dunn (1967) demonstrated that with massive population variance differences (e.g., 10:1 on all variables), sample size equality does not uniformly produce a robust test with samples smaller than about 50, although the number of variates is relevant, since situations that involve two or three variates show robustness at ns of 25, but those that involve 10 do not become robust until about $n = 100$. For $n_1 \neq n_2$, however, the test moves rapidly towards unacceptable Type I error rates as the degree of population covariance matrix heterogeneity is increased. Even for relatively mild between-populations dispersion differences and particularly as p becomes larger, a sample size ratio of even 2:1 produces unacceptable actual α values. With greater heterogeneity, the 2:1 $n_1:n_2$ ratio yields values that are completely out of hand, which is also true for the case of mild heterogeneity but more severely unequal ns (5:1). Also important is the fact that these results appear to be independent of sample size. In summary, it is clear that the T^2 procedure is not robust in the face of covariance matrix heterogeneity coupled with unequal ns, even for relatively mild departures from equality of the covariance matrices, sample sizes, or both.

Implications for Proper Use of the T^2 Procedure

It seems clear that prior to the T^2 analysis, the sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ should be inspected and inferentially tested for equality. The homogeneity issue has implications for not only the T^2 test but also for discriminant analysis—often employed to supplement T^2 results—in which heterogeneous covariance matrices, in the two-group case, suggest a quadratic rather than a linear function (Rao, 1965, pp. 488–489).

Statistical tests of the hypothesis of equal covariance matrices have been available for many years, with Bartlett's (1947) modification of the likelihood ratio criterion and Box's (1949) improved chi-square and F approximations (see Harris, 1975, pp. 85–86). These procedures, however, have been shown to be extremely sensitive to multivariate nonnormality (Hopkins & Clay, 1963; Mardia, 1971). Although little is known about it, a newer procedure due to Layard (see Timm, 1975, pp. 252–253) may be more robust in this regard.

Clearly, the user of the T^2 procedure is almost certainly on safe ground if the two samples have equal ns. There is little chance that a Type II error in the homogeneity test will cause a seriously biased test of the central hypothesis. Greenstreet and Connor (1974) showed that even for small (and equal) ns, scale differences in the variables of from 2 to 3 were detectable with high probability. And as noted, only for huge variance differences (e.g., 10:1) is the T^2 test biased. Thus, re-

jection of the homogeneity hypothesis will probably not be serious with equal *n*s.

With equal *n*s and rejection of the equality-of-covariance matrix hypothesis, the question remains of whether in a given instance the T^2 test is likely to be biased. As demonstrated, an operationalization of heterogeneity is the magnitude of the latent roots of the product $\Sigma_2\Sigma_1^{-1}$. Given a strict rescaling of the variables, these elements are the variance ratios of the variables. For sample data, therefore, if $\mathbf{D}$ with diagonal elements d_{ii}, is the diagonal matrix of latent roots of $S_2S_1^{-1}$ (such that $\Pi_i^p d_{ii} > 1$), then $(\Pi_i^p d_{ii})^{1/p}$ is a sample estimate approximately equal to a squared scaling factor, or an approximate overall estimated variance ratio. If this value is less than 5, for example—as will generally be true of real-world psychological data—then the user can safely ignore the inequality of the covariance matrices and proceed with the T^2 analysis. For larger *n*s—50 or beyond, for example—the T^2 will be sufficiently robust with heterogeneity values (as noted) of up to 10. Thus, the user can ascertain whether rejection of the homogeneity hypothesis is relevant for the T^2 analysis; as demonstrated, such a rejection will usually be irrelevant.

It is when $n_1 \neq n_2$ that problems arise. If the test of covariance matrix homogeneity is nonsignificant, no problem exists; but if this test is significant, the user is faced with the multivariate extension of the Behrens-Fisher problem. The following strategy—in the order of steps listed—seems to be a reasonable approach.

1. Ascertain whether one is in the positive or negative condition. A direct assessment of this comes from comparison of the determinants (understood as generalized variances) of S_1 and S_2. If either $n_1 > n_2$ and $|S_1| > |S_2|$ or $n_1 < n_2$ and $|S_1| < |S_2|$, we have the positive condition, whereas if the opposite obtains, we are in the negative condition. It is, of course, possible for $|S_1|$ to be equal to $|S_2|$, although $S_1 \neq S_2$. In such a case, however, the above taxonomy does not apply, and the effects of such heterogeneity may not be serious.

2. Given the positive condition, the T^2 test will be conservative. Thus the user should run the test and, if it is significant, reject the null hypothesis a fortiori.

3. Given the negative condition, the T^2 test will be liberal. Thus the user should run the test and, if it is nonsignificant, retain the null hypothesis.

4. If in the positive condition, T^2 is nonsignificant or if in the negative condition, T^2 is significant, two possibilities exist: (a) If the *n*s are not extremely different, they can be equalized by random deletion of subjects from the larger group. If in fact the null hypothesis is false, the loss of power may not be too great, and the previously significant T^2 (negative condition) may still be significant. The previously nonsignificant T^2 (positive condition) may now be significant. (b) If the *n*s are substantially different so that equalization would result in a massive loss of power or if the equalization performed when *n*s are not extremely different results in nonsignificant results, the user can employ one of several solutions to the multivariate Behrens-Fisher problem—reasonably precise approximations that do not require deletion of subjects. Such solutions can be found in an article by Ito (1969).

References

Anderson, T. W. *An introduction to multivariate statistical analysis*. New York: Wiley, 1958.

Bartlett, M. S. Multivariate analysis. *Journal of the Royal Statistical Society Supplement, Series B*, 1947, *9*, 176–197.

Box, G. E. P. A general distribution theory for a class of likelihood criteria. *Biometrika*, 1949, *36*, 317–346.

Chase, G. R., & Bulgren, W. G. A Monte Carlo investigation of the robustness of T^2. *Journal of the American Statistical Association*, 1971, *66*, 499–502.

Greenstreet, R. L., & Connor, R. J. Power of tests for equality of covariance matrices. *Technometrics*, 1974, *16*, 27–30.

Harris, R. J. *A primer of multivariate statistics*. New York: Academic Press, 1975.

Holloway, L. N., & Dunn, O. J. The robustness of Hotelling's T^2. *Journal of the American Statistical Association*, 1967, *62*, 124–136.

Hopkins, J. W., & Clay, P. P. F. Some empirical distributions of bivariate T^2 and homoscedasticity criterion *M* under unequal variance and leptokurtosis. *Journal of the American Statistical Association*, 1963, *58*, 1048–1053.

Hotelling, H. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 1931, *2*, 360–378.

Ito, K. On the effect of heteroscedasticity and nonnormality upon some multivariate test procedures.

In P. R. Krishnaiah (Ed.), *Multivariate analysis—II*. New York: Academic Press, 1969.

Ito, K., & Schull, W. J. On the robustness of the T^2_0 test in multivariate analysis of variance when variance–covariance matrices are not equal. *Biometrika*, 1964, *51*, 71–82.

Knuth, D. E. *The art of computer programming Vol. 2. Semi-numerical algorithms*. Reading, Mass.: Addison-Wesley, 1968.

Mardia, K. V. The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, 1971, *58*, 105–121.

Morrison, D. F. *Multivariate statistical methods* (2nd ed.). New York: McGraw-Hill, 1976.

Olson, C. L. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 1974, *69*, 894–908.

Pillai, K. C. S., & Sudjana. Exact robustness studies of tests of two multivariate hypotheses based on four criteria and their distribution problems under violations. *The Annals of Statistics*, 1975, *3*, 617–636.

Rao, C. R. *Linear statistical inference and its applications*. New York: Wiley, 1965.

Tatsuoka, M. M. *Multivariate analysis: Techniques for educational and psychological research*. New York: Wiley, 1971.

Timm, N. H. *Multivariate analysis with applications in education and psychology*. Monterey, Calif.: Brooks/Cole, 1975.

Received June 16, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1264–1271

Cerebral Electrotherapy: Methodological Problems in Assessing Its Therapeutic Effectiveness

Carmen L. von Richthofen
Memorial University of Newfoundland, St. John's, Newfoundland, Canada

Clive S. Mellor
Department of Psychiatry, Memorial University of Newfoundland, St. John's, Newfoundland, Canada

Cerebral electrotherapy (CET) appears to offer a safe and comfortable method of treating a variety of conditions, principally anxiety, depression, and insomnia. Attempts to assess its therapeutic efficacy have yielded widely differing results. This may be attributable to considerable variation in the electrical characteristics of the apparatus used, the duration of the treatment, and the placement of the electrodes. In some double-blind studies, the placebo condition has differed significantly from the treatment, and in most the ideal situation with a blind machine operator, subject, and assessor, has not been achieved. Until the methodology for assessment improves and the treatment procedure is standardized, it will be impossible to determine if CET is an effective treatment, and if it is, whether its mode of action is attributable to a direct effect on the brain or to relaxation, suggestion, or tactile stimulation.

Reports from Eastern Europe and particularly the Soviet Union suggest that cerebral electrotherapy (CET) or electrosleep treatment, holds considerable promise in the treatment of a variety of conditions. This review examines the methodological problems that have attended controlled studies that attempt to validate these claims. First, an overview of what the term CET implies is provided.

Overview

Electrosleep, or cerebral electrotherapy (CET), is a somatic therapy characterized by the passage of a low-amplitude, pulsating direct electrical current around and through the cranium. Originally, electrosleep was intended to induce a state of natural sleep, "a state of consciousness grossly indistinguishable from ordinary sleep, produced by the direct action of a weak rhythmic current on the brain of a cooperative subject in a nondistracting environment" (Boblitt, 1969, p. 9).

Requests for reprints should be sent to Clive S. Mellor, who is also at the St. Clare's Mercy Hospital, St. John's, Newfoundland, Canada A1C 5B8.

Pavlov is said to have provided the concept and rationale for electrically produced sleep therapy (Boblitt, 1969; Obrosow, 1959). His concept of cerebral protective inhibition was based on the idea that a prolonged, monotonous, weak stimulus, such as a mild pulsating electric current, which is applied to the central nervous system under conditions of comfort, allows the brain cells to rest and permits restoration of function.

Investigations into the effects of direct electrical currents have been going on since the 19th century. Included in the techniques was that of electroanesthesia, or the depression of consciousness, a state determined by the basic criterion of nonresponsiveness to pain and achieved by means of an electric current usually in the range of 20 Hz to 1 KHz (Brown, 1975). A second technique, polarization, resembles and is often confused with CET. The major differences involve the use of constant rather than pulsating current and the positioning of electrodes on the arm or leg rather than solely on the cranium. Polarization is thought to be useful in producing mood changes (Lippold & Redfearn, 1964).

The modern view of electrosleep therapy as a technique distinct from electroanesthesia

began with the Soviet researcher Giljarowskii in the early 1950s (Lewis, 1966).

The view that the sole purpose of electrosleep therapy was to induce a sleeplike state to promote functional recovery of cerebral cells influenced the type of research carried out in the Soviet Union and Europe. Most of this research has been presented at the International Symposia for Electrosleep and Electroanesthesia, first held in 1966, with subsequent meetings in 1969 and 1972. It was after the first symposium that the notion of a therapeutic, protective, artificial sleep was gradually replaced by the idea that the direct action of the current itself was the healing force:

> Our experience has shown that sleep in the course of the individual session is not an absolute condition of success of therapy . . . . The passage of the pulse current through the brain is of greater importance for a curative effect than the achievement of the condition of sleep in one course of treatment. (Van Poznak, 1969, p. 507)

Accordingly, Wageneder proposed adopting the term cerebral electrotherapy (CET) to replace electrosleep in that it more accurately reflected the type of treatment involved (Wageneder & St. Schuy, 1970).

It was after the first International Symposium for Electrosleep and Electroanesthesia in 1966 that North American researchers became interested in CET. Before this symposium, most of the work on CET had been done in the Soviet Union and Europe. Translations of these works revealed sweeping claims for the beneficial effects of CET in a wide variety of disorders in the fields of psychiatry, surgery, dermatology, obstetrics, and pediatrics, but experimental controls were inadequate or nonexistent (Van Poznak, 1969). The American interest in CET research marked the beginning of a somewhat more objective assessment of its effects.

Whereas the Soviet and European researchers had presented a united front in their favorable opinion of the beneficial effects of CET, the American investigators were markedly divided in their opinions, in spite of the fact that the focus of CET research in America had been narrowed down to the main target disorders of anxiety, depression, and insomnia and to certain physiological effects. In general, this division of opinion holds among authors of uncontrolled studies as well as among those who used double-blind procedures in their investigations.

The difficulty in drawing any conclusions is compounded by the considerable differences in the research methods so that comparisons between studies are impossible. This is illustrated by examining in detail eight recent studies that were more rigorous than others because they at least used some kind of double-blind experimental procedure.

Techniques of CET

The technique of administering CET encompasses a bewildering array of different procedures, particularly when the type of electrical stimulation, the duration of the treatment, and the placement of the electrodes are varied.

Electrical Parameters

Either a DC or an AC, in which the current flow regularly changes direction, has been used. Some researchers have added to these a DC bias (Itil, Gannon, Akpinar, & Hsu, 1971; Marshall & Izard, 1974). DC is most frequently used and adheres to the original electrosleep technique and theory in that a constant direction of current is thought to be essential for ensuring unidirectional cerebral cell reactions, an important factor in producing desired physiological modifications (Obrosow, 1959). The pulse frequency, measured in impulses per second or Hz, may vary from 30 Hz to 100 Hz (Rosenthal, 1972a), with a pulse width ranging from 1 to 2 msec (Straus, Elkind, & Bodian, 1964). The supply voltage may vary from 10 to 20 V. The current amplitude at which tingling is usually felt ranges from .1 to .5 mA. These current parameters depend on the type of CET device used, and no two American-made models seem to possess the same electrical characteristics (Brown, 1975). However, the combination of 100 Hz with a pulse duration of 2 msec and an amplitude of up to 1.5 mA are the most frequently used parameters in double-blind studies.

Table 1
Eight Double-Blind Studies and the Technique of Cerebral Electrotherapy

Authors	Current	Frequency (in Hz)	Pulse length (in msec)	Electrode placement	Treatment session No.	Treatment session Duration (in minutes)
Straus, Elkind, & Bodian (1964)	DC	30–40	1.8–2.0	Eyelid +, mastoid −	6–12	30
Rosenthal (1972a)	DC	100	1.0	Mastoid +, orbit −	5	30
Weiss (1973)	DC	45–240	.4–1.2	Brow, nape of neck[a]	24	1–5 2–10 21–15[b]
Feighner, Brown, & Oliver (1973)	DC	100	1.0	Eyelids, mastoids[a]	20	30
Tomsovic & Edwards (1973)	AC	100	2.0	Orbits, mastoids[a]	5	30
Hearst, Cloninger, Crews, & Cadoret (1974)	AC, DC	100	2.0	Mastoids +, brow −	5	30
Marshall & Izard (1974)	DC	100	1.0	Eyelid +, mastoid −	5	30
Moore, Mellor, Standage, & Strong (1975)	DC	100	2.0	Not given	10	30

[a] The polarities of the electrode placements are not given.
[b] The duration of treatment was increased gradually.

Duration of Treatment

Treatment sessions have been known to vary from 30 minutes for 5 consecutive days (Rosenthal, 1972a) to 2 hours daily for a period of several months (Wageneder, Iwanovsky, & Dodge, 1969). It has been reported that exposure to CET for more than 2 hours per session resulted in morning dizziness and a degree of unsteadiness in walking that lasted for a few hours (Iwanovsky & Dodge, 1968). The average duration of exposure time in most of the important studies was 30 minutes per session over an average of 10 sessions.

Electrode Placement

The felt pads attached to the electrodes are either soaked in water (Weiss, 1973) or a saline solution (Feighner, Brown, & Olivier, 1973) or prepared with saline paste (Hearst, Cloninger, Crews, & Cadoret, 1974). A pair of electrodes is placed either directly on the eyelids (Lewis, 1966) or the brow (Brown, 1975). A second pair is usually placed over the mastoids. The forehead electrodes are negatively charged cathodes and those at the mastoids are positively charged anodes (Rosenthal, 1972a), but in some cases their placement is reversed (Marshall & Izard, 1974). There is no firmly established rule about polarity, but the placement of negatively charged electrodes on the forehead and positively charged electrodes over the mastoids is more frequent in the literature (Iwanovsky & Dodge, 1968). However, the direction of current flow appears to be an important factor according to one of the original users of CET, Giljarowskii, who felt that current entering through the orbital fissure and leaving from the mastoid processes was the best method of ensuring electrical penetration of the brain (Boblitt, 1969, p. 12). Since current conventionally flows from the positive to the negative electrodes, the anodes should be anterior relative to the cathodes. A significant reason for such a placement of electrodes is that the only investigation that has demonstrated intracerebral current flow (Dymond, Coger, & Serafetinides, 1975) followed this procedure. The differences in technique between the eight double-blind studies already referred to are set out in Table 1.

Table 2
Eight Cerebral Electrotherapy Double-Blind Studies: Variables in Experimental Design and Outcome

Authors	Target condition	No. of subjects	Cross-over design	Ideal double blind	Tingling sensation during treatment		Experientially identical[a]	Significant difference[a]
					Active	Placebo		
Straus, Elkind, & Bodian (1964)	Insomnia	34	No	No	Yes	Only initially	No	Yes
Rosenthal (1972a)	Anxiety, depression, insomnia	22	No	No	Yes (with loud noise)	No (with loud noise)	No	Yes
Weiss (1973)	Insomnia	10	No	Yes	Yes	Only initially	No	Yes
Feighner, Brown, & Oliver (1973)	Anxiety, depression, insomnia	23	Yes	No	Yes	Only initially	No	Yes
Tomsovic & Edwards (1973)	Tension, head/stomach pain	43	No	No	Yes	Only initially	No	No
Hearst, Cloninger, Crews, & Cadoret (1974)	Anxiety, depression, insomnia	28	No	Yes	Only initially	Only initially	Yes	No
Marshall & Izard (1974)	Depression	40	No	No	Yes	Yes	Yes	No
Moore, Mellor, Standage, & Strong (1975)	Anxiety, depression, insomnia	10	Yes	No	Only initially	Only initially	Yes	No

[a] For comparison of active and placebo treatements.

Setting for Treatment and Method of Administration

It is generally agreed that the best treatment setting is a quiet darkened room in which the patient lies comfortably on a bed. After the electrodes are applied, the current is turned on, and the amplitude is increased slowly to avoid any unpleasant sensation, until the level is reached at which a slight tingling sensation occurs. Some clinicians then reduce the current to a level at which the patient experiences no cutaneous sensation (see Table 2).

Complications and Contraindications

CET is an attractive treatment because it is said to have no cumulative side effects (Weiss, 1973), complications, or contraindications, to be nontoxic and usable with drugs and other therapies, and to be a simple procedure to carry out (Weinberg, 1969). However, side effects such as blurring of vision, thought to be the result of electrode pressure on the eye, and dizziness as well as slight burns on the skin at the electrode sites have been reported (Frankel, 1974; Koegler, Hicks, & Barger, 1971; Rosenthal & Wulfsohn, 1970). Moreover, contraindications such as epilepsy, blood diseases, malignant tumours, cerebrovascular disorders, and heart disease (Chumakova & Kirillova, 1976) as well as various forms of psychosis (Rosenthal, 1972b) have been reported. All of the double-blind studies excluded such patients.

Mode of Action of CET

Before double-blind designs can be discussed, the theories of the mode of action of CET should be considered. These theories must be taken into account when the qualifications of the different placebo conditions are reviewed.

There are two main schools of thought about the action of CET: those researchers who postulate that it has a direct effect on the brain and those researchers who attribute its effect to other causes.

Direct Effect

The effects of CET treatment are attributed to the direct action of the current on the cerebral cells. Rush and Driscoll (1968), in their work on a theoretical model of current flow in the human head, assumed that the current entered the cranium via the frontal surface electrodes. Their calculations suggested that 45% of the electrical output actually entered the brain. Those who favor the direct effect theory hold that the current traversing the brain induces protective inhibition that creates favorable recovery of cerebral cells along with sedation and normalization of the central nervous system (CNS) processes (Banshchikov, 1967; Brand, 1970).

One criticism of the direct effect theory is that the current is too weak to pass from the skin through the skull and other tissues to effect changes in the brain. However, Dymond et al. (1975) provided direct evidence that CET is capable of producing electrical changes in the brain. The proponents of the direct effect theory have mainly been the Soviet and European researchers who were influenced by the original Pavlovian concept of protective inhibition. Kalinowsky (1969) suggested that the therapeutic usefulness of CET lay in the "rhythmic nature of a peripheral stimulation" (p. 172). Iwanovsky and Dodge (1968) described a Soviet work in which the author conclusively stated that CET was a rhythm therapy in which the electrical current provided a kind of "electromassage" that polarized cells and normalized tissue metabolism.

Indirect Effect

The effects of CET treatment are due to the indirect action of the current. The normalization of only the peripheral autonomic elements of the nervous system are involved, and the effect on the CNS is a secondary one involving a variety of mechanisms (Iwanovsky and Dodge, 1968).

The first mechanism is thought to be relaxation, attributable to lying down in a quiet, comfortable, semidark setting. Some reviewers of CET literature have repeatedly stressed the importance of the elements of suggestion inherent in a procedure that calls for patients to lie comfortably in a quiet darkened room and submit to a treamtment that they are told will relax them (Boblitt, 1969; Frankel, 1974; Lewis, 1966).

A second mechanism is sensory stimulation, whereby the rhythmic cutaneous sensations experienced in treatment may alone account for the clinical effects. Such sensory stimuli have been found to induce sleep (Lovell & Morgan, 1942; Oswald, 1960).

A third mechanism is suggestion or a placebo effect, whereby patients who are referred by physicians to undergo a special type of therapy involving neither drugs nor psychotherapy are under the impression that this represents a new type of cure for their specific ailment. The powerful suggestion inherent in electrical apparatus applied to healing has been recognized since the time of Mesmer. The notion that a possible placebo effect played an unimportant role in CET research was promulgated by Giljarowskii, a notion which according to Frankel (1974) probably encouraged many subsequent investigators to adopt a less rigorous approach to their experiment.

Double-Blind Studies

The ideal double-blind procedure is one in which the subject, the operator of the machine, and the individual assessing the effects are all blind. The methodological problems that occur when this is attempted are considerable. Only two of the eight studies listed in Table 2 were able to meet this ideal double-blind or more accurately, triple-blind condition (Hearst et at., 1974; Weiss, 1973).

Treatment Conditions

In three studies the active and placebo CET conditions were identical (Hearst et al., 1974; Marshall & Izard, 1974; Moore, Mellor, Standage, & Strong, 1975). In all of these no statistically significant differences were found between active and placebo treatment (see Table 2). Two types of identical active and placebo CET conditions were devised. One method was to lower the current to just below the point at which tingling was perceived in active treatment. This would be done after subjects had been allowed to experience an initial tingling sensation. Thus an active treatment condition experientially identical to placebo CET, in which the current was turned off completely, was produced.

An alternative method was that of Marshall and Izard (1974), which used positive and negative frontal electrodes and thus enabled the subjects to experience the cutaneous sensation. There was no active treatment because the current was flowing just through the inch of skin separating the electrodes. The electrodes were applied to the mastoid as usual but were not connected to the current. Although all of the studies purported to have kept subjects blind, the first five studies listed in Table 2 did not create active and placebo treatment conditions that were experientially identical to subjects. In some of these studies, subjects felt a tingling sensation throughout active treatment but experienced it only for a brief period at the beginning of placebo treatment. In the placebo treatment the machine would be turned off, persumably without the subjects' knowledge. In other studies, a noise or light used in both treatment conditions was supposed to represent the giving of CET, although the tingling sensation felt in active treatment was absent during placebo treatment. All but one of these five investigations found a significant difference between the active and placebo treatment. The results of these eight studies suggest an association between identical peripheral stimulation and negative outcome and different peripheral stimulation and positive outcome, when active treatment is compared with placebo, although it does not reach statistical significance ($p = .07$, Fisher's exact test). Ensuring experimentially identical active and placebo treatment conditions may be one of the most important factors when the outcome of CET is being assessed.

The implications of these results can be related to the two theories of CET mode of action. Two studies in which subjects felt tingling in neither active nor placebo treatment possibly represented a test of the direct effect theory: The intervening variable of peripheral or rhythmic sensation was not present; therefore, the direct effect alone was being tested. No statistically significant difference between active and placebo treatment was found. In a third study, in which subjects experienced tingling in both active and placebo treatment,

peripheral sensation was held constant. Although both active and placebo treatment groups improved, there was no significant difference between group improvement levels. Thus the indirect effect theory would appear to have been supported once more.

Machine Operator

An additional aspect to ensuring that subjects are blind, given experientially identical active and placebo conditions, is that the operator of the machine also be blind. Frankel (1974) has pointed out the possibility of either indirect verbal or nonverbal communications between the operator and the subject. The application of active or placebo treatment may be identified by the subject from such subtle communications. Such procedures require an intermediary who would only operate the machine while another individual would only interact with subjects. An ideal double-blind procedure is probably best achieved with the use of specially built machines with hidden switches (Frankel, 1974; Weiss, 1973). Perhaps because of the need for extra staff and equipment, only two of the eight double-blind studies were carried out under these conditions (Hearst et al., 1974; Weiss, 1973). In the Weiss study, significant differences between active and placebo CET were found; in the latter study no significant differences were noted.

Assessment of Effects

To prevent the influence of knowledge of type of treatment on assessment, objective psychological measures such as inventories completed by subjects or rating scales completed by a blind clinical assessor should be used. All of the eight double-blind studies under discussion complied with one or both of these conditions.

All these studies used relatively small numbers of subjects. This raises the problem of substantial differences between treatment and control groups before the treatment is administered. To some extent this problem can be obviated by using a crossover design like that employed by Feighner et al. (1973) and Moore et al. (1975). An important aspect to assessment of CET effects, and one that has thus far not been explored in conjunction with psychological and clinical effects, is that of physiological effects. There is a body of research on CET that deals exclusively with physiological effects of CET and does not concern itself with psychological or clinical effects. Probably the most thorough method of properly evaluating the therapeutic effectiveness of CET would be to design a study in which CET's effect on a target disorder is assessed with objective psychological, physiological, and clinical measures and in which subjects serve as their own controls in a crossover design. Such a study has been completed by the authors and will be published in the near future.

Conclusion

It is clear that the confusion about the actual therapeutic value of CET stems in part from the diversity of methodological approaches and from the varying degrees of scientific rigor that have been used in CET research.

Any further research on CET would be of little value unless the scientifically sound components of all these studies are combined to produce a design that allows no room for bias.

Suggestions for Future Investigations of CET

Based on this discussion of the various techniques and methodologies of CET research, the designers of future studies should consider the following points:

1. Electrical parameters have been limited to a pulse frequency of 100 Hz and a duration of 2 msec. Little consideration has so far been given to the effects of varying the pulse frequency and duration.
2. Treatment duration—The effects of the duration and frequency of treatments have not been systematically studied.
3. Electrode placement and polarity.
4. Treatment conditions—Ideally, the subjects, the operator of the machine, and the clinical assessors should be blind. Subjects' phenomenological experience of treatment

should be identical under all experimental conditions.

5. Assessment—Objective psychological and physiological measures should be used, along with blind assessment of the changes in the subjects' clinical condition.

References

Banshchikov, V. M. Present status of electrosleep in the USSR. In F. M. Wageneder & G. St. Schuy (Eds.), *Electrotherapeutic sleep and electroanesthesia*. Amsterdam: Excerpta Medica Foundation, 1967.

Boblitt, W. E. Electrosleep as a sleep induction method. *The Psychiatric Forum*, 1969, *1*, 9–14.

Brand, J. Electrosleep therapy for migraine and headache. In F. M. Wageneder & G. St. Schuy (Eds.), *Electrotherapeutic sleep and electroanesthesia*. Amsterdam: Excerpta Medica Foundation, 1970.

Brown, C. C. Electroanesthesia and electrosleep. *American Psychologist*, 1975, *30*, 402–410.

Chumakova, L. T., & Kirillova, Z. A. Electrosleep as an effective outpatient treatment for nervous and psychological disorders. In R. M. Suinn & R. D. Weigel (Eds.), *Innovative medical psychiatric therapies*. Baltimore, Md.: University Park Press, 1976.

Dymond, A. M., Coger, R. W., & Serafetinides, E. A. Intracerebral current levels in man during electrosleep therapy. *Biological Psychiatry*, 1975, *10*, 101–104.

Feighner, J. P., Brown, S. L., & Olivier, J. E. Electrosleep therapy. *Journal of Nervous and Mental Disease*, 1973, *157*, 121–128.

Frankel, B. L. Research on cerebral electrotherapy (electrosleep): Some suggestions. *American Journal of Psychiatry*, 1974, *131*, 95–98.

Hearst, E. D., Cloninger, C. R., Crews, E. L., & Cadoret, R. J. Electrosleep therapy: A double-blind trial. *Archives of General Psychiatry*, 1974, *30*, 463–466.

Itil, T., Gannon, P., Akpinar, S., & Hsu, W. Quantitative EEG analysis of electrosleep using frequency analyzer and digital computer methods. *Electroencephalography and Clinical Neurophysiology*, 1971, *31*, 294.

Iwanovsky, A., & Dodge, C. H. Electrosleep and electroanesthesia: Theory and clinical experience. *Foreign Science Bulletin*, 1968, *4*, 1–64.

Kalinowsky, L. B., & Hippius, H. *Pharmacological, convulsive and other somatic treatments in psychiatry*. New York: Grune & Stratton, 1969.

Koegler, R. R., Hicks, S. M., & Barger, J. H. Medical and psychiatric use of electrosleep: Transcerebral electrotherapy. *Diseases of the Nervous System*, 1971, *32*, 100–104.

Lewis, J. A. In R. L. Williams & W. B. Webb (Eds.), *Sleep therapy: A bibliography and commentary*. Springfield, Ill.: Charles C Thomas, 1966.

Lippold, O. C. J., & Redfearn, J. W. T. Mental changes resulting from the passage of small direct currents through the human brain. *British Journal of Psychiatry*, 1964, *110*, 768–772.

Lovell, G. D., & Morgan, J. J. B. Physiological and motor responses to a regular recurring sound: A study in monotony. *Journal of Experimental Psychology*, 1942, *30*, 435–451.

Marshall, A. G., & Izard, C. C. Cerebral electrotherapeutic treatment of depressions. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 93–97.

Moore, J. A., Mellor, C. S., Standage, K. F., & Strong, H. A double-blind study of electrosleep for anxiety and insomnia. *Biological Psychiatry*, 1975, *10*, 59–63.

Obrosow, A. E. Electrosleep therapy. In E. Licht (Ed.), *Therapeutic electricity and ultraviolet radiation*. New Haven, Conn.: Editor, 1959.

Oswald, I. Falling asleep open-eyed during intense rhythmic stimulation. *British Medical Journal*, 1960, *1*, 1450–1451.

Rosenthal, S. H. Electrosleep: A double-blind clinical study. *Biological Psychiatry*, 1972, *4*, 179–185. (a)

Rosenthal, S. H. Electrosleep therapy. In J. H. Masserman (Ed.), *Current psychiatric therapies* (Vol. 12). New York: Grune & Stratton, 1972. (b)

Rosenthal, S. H., & Wulfsohn, N. L. Electrosleep. *Journal of Nervous and Mental Disease*, 1970, *151*, 146–151.

Rush, S., & Driscoll, D. A. Current distribution in the brain from surface electrodes. *Anesthesia and Analgesia*, 1968, *47*, 717–723.

Straus, B., Elkind, A., & Bodian, C. A. Electrical induction of sleep. *American Journal of Medical Science*, 1964, *248*, 514–520.

Tomsovic, M., & Edwards, R. V. Cerebral electrotherapy for tension-related symptoms in alcoholics. *Quarterly Journal of Studies on Alcohol*, 1973, *34*, 1352–1355.

Van Poznak, A. Advances in electrosleep and electroanesthesia during the past decade. *Clinical Anesthesia*, 1969, *3*, 501–520.

Wageneder, F. M., Iwanovsky, A., & Dodge, C. H. Electrosleep (cerebral electrotherapy) and electroanesthesia: The international effort at evaluation. *Foreign Science Bulletin*, 1969, *5*, 1–104.

Wageneder, F. M., & St. Schuy, G. (Eds.). *Electrotherapeutic sleep and electroanesthesia*. Amsterdam: Excerpta Medica Foundation, 1970.

Weinberg, A. Clinical observations in the use of electrosleep. *Journal of American Society of Psychosomatic Dentistry and Medicine*, 1969, *16*, 35–39.

Weiss, M. F. The treatment of insomnia through the use of electrosleep: An EEG study. *Jounral of Nervous and Mental Disease*, 1973, *157*, 108–120.

Received June 23, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1272–1279

Temporal Versus Spatial Information Processing Theories of Hippocampal Function

Paul R. Solomon
Williams College

A series of articles by Black, Nadel, O'Keefe, and their co-workers propose that the primary function of the hippocampus is to process spatial information. Although the spatial information processing view of hippocampal function accounts for much of the available data, it cannot account for the data from the classically conditioned rabbit nictitating membrane response preparation. The present article reviews these data and suggests that the hippocampus is involved in the processing of temporal as well as spatial information.

A number of recent studies by Black, Nadel, O'Keefe, and their co-workers propose that the primary role of the hippocampus is to process spatial information (see Nadel & O'Keefe, 1974; Nadel, O'Keefe, & Black, 1975, for reviews). According to this theory, the hippocampus acts as part of a neural system that forms a cognitive map of the environment. Central representations of separate places in the environment as well as the relationship of one place to any other place are represented in this system. Thus once an animal has located itself in the environment by using the available cues, it can use its spatial mapping system to locate other places.

Central to the spatial hypothesis of hippocampal function is that the role of the hippocampus is to process spatial but not temporal information. As O'Keefe and Black (1979) point out, cues are only critical to the extent that they allow the animal to identify a starting point (e.g., the start box in a maze). Once this point has been identified, the animal can find any place in the environment by using the mapping system. Furthermore, once the starting point is perceived, the animal's spatial map is not sensitive with respect to its own body position or orientation, nor is the map sensitive to cues in the environment (O'Keefe, 1976).

To support the spatial information processing view of hippocampal function, its proponents have presented data from both electrophysiological (O'Keefe, 1976; O'Keefe & Black, 1979; O'Keefe & Dostrovsky, 1971; Olton, Branch, & Best, 1978) and lesion (Black, Nadel, & O'Keefe, 1977; O'Keefe & Black, 1979; O'Keefe, Nadel, Kieghtly, & Kill, 1975; Olton, Walker, & Gage, 1978) studies. The data and logic from both these lines of research strongly implicate the hippocampus in spatial mapping. The authors, however, overlook an accumulating body of literature on the role of the hippocampus in aversive classical conditioning of the rabbit's nictitating membrane response (NMR). In this article I argue that place learning plays little if any role in classical conditioning of the rabbit's NMR and if this is the case, the data pertaining to hippocampal function in this preparation cannot be explained by the spatial information processing hypothesis.

The Spatial Information-Processing Hypothesis and Classical Conditioning

The proponents of the spatial information-processing view of hippocampal function do not

Some of the research described in this proposal was supported by National Science Foundation Grant BNS-77-14871 and a faculty research grant.

I would like to thank Andrew Crider, John W. Moore, W. Ronald Salafia, and Richard F. Thompson for their helpful comments on an earlier version of this article.

Requests for reprints should be sent to Paul R. Solomon, Department of Psychology, Bronfman Science Center, Williams College, Williamstown, Massachusetts 01267.

directly address the question of what changes in aversively motivated classically conditioned behaviors might be expected in animals with hippocampal lesions. They do, however, point out that place learning does not involve learning about the temporal relationships between stimuli:

> When an animal is learning about the spatial relationships among stimuli, we assume that it employs place strategies. When it is learning about the temporal relationship between stimuli we assume that it employs cue strategies. (Black et al., 1977, p. 1108.)

The authors assume that place learning involves the acquisition of knowledge of the spatial relationships between two stimuli and that this knowledge is acquired independently of (and via different neural substrates than) knowledge of temporal relationships (cue learning). The hippocampus, in their view, is critical for learning the spatial aspects of a relationship between stimuli but unimportant for learning the cuing of temporal relationships. Black et al. (1977) make this explicit in reviewing the data that indicate no deficit in taste aversion learning in animals with hippocampal lesions: "This result is to be expected, since this procedure (taste aversion) does not seem to involve spatial strategies, at least as it has been employed so far" (p. 1123).

The strong implication here is that hippocampal lesions should not affect the temporal associations made in a classical conditioning preparation such as the rabbit NMR. As I point out in subsequent sections, although there are some tasks in which hippocampal lesions do not seem to affect the classically conditioned NMR, there are others in which animals with hippocampal lesions differ from controls. Furthermore, the electrophysiological data and the data from brain stimulation studies also implicate the hippocampus in classical conditioning of the NMR.

Absence of Spatial Cues in the Rabbit NMR Preparation

Black et al. (1977, p. 1108) state that when an animal is learning spatial relationships between stimuli, it is using place strategies, and when it is learning temporal relationships between stimuli, it is using cue strategies. The authors also point out that, although an animal performs successfully in a particular situation, it may be difficult to determine to what degree each strategy is being used. Consequently, it may be necessary to make a priori assumptions about each task.

Although it is possible that spatial learning is involved in classical conditioning paradigms such as the conditioned emotional response (CER) (see Black et al., 1977, p. 1114), it is unlikely that spatial cues play any role in the rabbit NMR preparation. Here, the animal remains virtually motionless throughout the conditioning session, and the conditioned stimuli (CS) and unconditioned stimuli (UCS) are delivered in the same spatial locations at all times. Thus any associations, for example, between the CS and the UCS, would necessarily be temporal in nature.

The rabbit NMR preparation used in most laboratories is a variation of that described by Gormezano (1966). The rabbit is restrained in a Plexiglas box with an adjustable plate and ear clamp securing the head and a second plate placed over the animal's back to restrict body movement. Some experimenters further restrain the animal's head by implanting long bolts in the animal's skull and fastening them to the plexiglas box (e.g., Frey, Maisiak, & Dugue, 1976; Mis, 1977). Animals are typically run in individual sound-attenuated and darkened chambers. A panel in front of each chamber contains two lights that serve as visual CSs and speakers for delivering auditory CSs. Several laboratories (see Wagner, Rudy, & Whitlow, 1973) also use a vibratory CS delivered to the back of the animal.

The UCS is typically a 1–3 mA infraorbital shock, although some experimenters prefer an air puff. The unconditioned response (UCR) and conditioned response (CR) are lateral movements of the nictitating membrane. This response is recorded by attaching the shaft of a potentiometer to the NM, thus transducing the response into a DC signal. A typical session in our laboratory consists of 100 CS–UCS pairings in a 50-minute session.

Lesion Studies

Since acquisition of the rabbit's NMR is not dependent on spatial strategies, the spatial information processing theory would predict

that hippocampal lesions should have no effect on this behavior. This, in fact, seems to be the case. Several studies (Schmaltz & Theios, 1972; Solomon, 1977; Solomon & Moore, 1975) reported no differences in acquisition of the NMR in animals with bilateral aspiration lesions of the dorsal hippocampus, animals with lesions of the overlying cortex, or unoperated controls. The dependent measure in these studies was either total CRs or trials to criterion. This, however, does not rule out the possibility that other more subtle aspects of the nacent CR, such as amplitude, latency, or interstimulus interval (ISI) shifts, could have been affected.[1] In addition, there are electrophysiological data (see the Electrophysiological Studies section) which suggest that the intact hippocampus is involved in acquisition of the CR.

Although the acquisition data may be consistent with the spatial processing view of hippocampal function, there are data from other tasks conducted in the rabbit NMR preparation that are not. Like simple acquisition, these tasks require that the animal learn a temporal relationship between a CS (or a series of CSs) and a UCS (or the absence of a UCS). Unlike acquisition, these tasks require that the animal learn not to respond in certain circumstances. In these situations the animal must learn whether a particular CS is relevant, that is, if it uniquely predicts the occurrence or nonoccurrence of the UCS. These tasks require the animal to learn temporal relationships between CSs and UCSs and thus, according to spatial information processing theory, should be performed equally well by animals with and without the hippocampus.

Latent inhibition (LI) is one behavior that is disrupted in animals with hippocampal lesions. In this paradigm the animal is preexposed to the to-be-conditioned CS, and this preexposure results in retarded acquisition of the CR when the CS is subsequently paired with the UCS in a conditioning paradigm. In one study conducted on the LI effect (Solomon & Moore, 1975), we reported that whereas 450 tone preexposures resulted in a decrement in conditioning for normal rabbits and for rabbits with cortical ablations, animals with dorsal hippocampal ablations showed no such decrement; that is, they conditioned as fast as nonpreexposed controls. Our interpretation of these data was that the hippocampus is part of a system involved in learning to ignore irrelevant stimuli. To further test this view of the role of the hippocampus, we investigated the effects of dorsal hippocampal ablations in Kamin's (1968, 1969) two-stage blocking paradigm.

The typical blocking paradigm in the rabbit NMR preparation consists of a two-group design (cf. Marchant & Moore, 1973). In Stage 1 the blocking group is presented with a tone that is paired with an eyeshock UCS until the CR is well established. Animals in the control condition are yoked to the blocking animals and simply sit for a corresponding amount of time with no CS or UCS presentations. In Stage 2 both groups are conditioned to a compound CS that consists of the tone from Stage 1 plus a light. After both groups display a high level of conditioning to the compound, the test phase is introduced. During testing, all animals are presented with nonreinforced presentations of the tone interspersed with nonreinforced light presentations. In general, whereas animals in the control condition give CRs to both the tone and light, animals in the blocking groups respond only to the tone (Marchant & Moore, 1973).

Although Kamin (1968) initially suggested that prior conditioning to the tone caused the animal not to notice the light when it was presented in compound with the tone, more recent evidence (Kamin, 1969; Mackintosh, 1973, 1975) suggests that the animal does initially attend to the redundant CS but does not condition to it, since the light provides no new information regarding the reinforcing event. Thus the paradigm is similar to LI in that the animal must learn to ignore an irrelevant stimulus. If a tuning-out process similar to the one in latent inhibition is operating here, hippocampal lesions should disrupt the blocking effect. Solomon (1977) reported just this finding. Although normal rabbits and rabbits with cortical lesions showed the typical blocking effect, hippocampectomized rabbits did not. Animals with dorsal hippocampal ablations re-

---

[1] I am grateful to Michael M. Patterson (Note 1) of Ohio University for suggesting this as yet untested possibility.

sponded to both the tone and light during testing.

Conditioned inhibition (CI) is yet another behavior in which an animal must learn not to respond in certain situations. Both we (Solomon, 1977; Moore, Mis, & Solomon, Note 2) and Black et al. (1977; Nadel et al., 1975) agree that the hippocampus is not critical to inhibitory processes. But whereas Black et al. argue against the inhibitory hypothesis of hippocampal function in favor of a spatial hypothesis, we maintain that CI, unlike blocking and LI, does not involve learning to ignore an irrelevant stimulus and thus should not be affected by hippocampal damage.

Although a number of theoretical accounts of hippocampal function have stated that the hippocampus plays a direct role in Pavlovian conditioned (e.g., Kimble, 1968) or internal (e.g., Douglas, 1967) inhibition, only recently has this been directly tested. In the Pavlovian CI paradigm, the animal is taught to discriminate between CS_A, which is always followed by the UCS, and a compound consisting of $CS_A + CS_B$, which is never followed by the UCS. Since CS_B is negatively correlated with the UCS, it should become inhibitory (Hearst, 1972; Rescorla, 1969). That is, when CS_B is presented alone and reinforced, acquisition of the CR should be retarded relative to controls. Experimenters using the rabbit NMR preparation have indicated that this is the case in normal rabbits (Marchant, Mis, & Moore, 1972; Marchant & Moore, 1974; Mis, 1977), and Solomon (1977) showed that this was also true for rabbits with dorsal hippocampal ablations.

Black et al. (1977) would likely argue that the failure to find a disruption of CI indicates that no spatial cues were involved. This appears to be true; however, this is not the factor that separates this paradigm from blocking and LI. Rather, we propose that the difference between CI, in which hippocampal lesions have no effect, and blocking and LI, in which they do, is that CI does not involve the tuning out of an irrelevant stimulus. In the Pavlovian CI paradigm, despite the fact that CS_B is not followed by the UCS, CS_B cannot be ignored. If the animal attempts to tune out CS_B, it could not make the distinction between CS_A and CS_{A+B}. Thus the failure to find a disruption of CI, at least from our theoretical perspective, was not surprising.

In summary, the data on the role of the hippocampus in acquisition of the classically conditioned NMR may be consistent with the spatial hypothesis of hippocampal function (but see the next section). Hippocampal lesions do, however, disrupt blocking and LI of the rabbit's NMR, but they do not seem to affect CI in this paradigm. Since each of these tasks involves temporal relationships between stimuli, it is not likely that a spatial hypothesis of hippocampal function can account for the differences. Rather, it appears that these data are best explained in terms of the role of the hippocampus in learning to ignore irrelevant stimuli.

Electrophysiological Studies

As I indicated earlier, much of the support for the spatial mapping hypothesis stems from electrophysiological data (O'Keefe & Black, 1979; O'Keefe, 1976; O'Keefe & Dostrovsky, 1971; Olton et al., 1978). These studies indicated that single-unit activity in the hippocampus was related to place but not to cue learning. There is, however, substantial electrophysiological data from the rabbit that implicates the hippocampus in classical conditioning of the rabbit's NMR. This work has been carried out by Thompson and his coworkers (Berger, Alger, & Thompson, 1976; Berger & Thompson, 1978a, 1978b; Berry & Thompson, 1978; Thompson, 1976).

Berger et al. (1976) reported increased neural activity in the hippocampus during pairings of a tone CS and corneal air-puff UCS. They found an increase in neural activity in both the pyramidal and granule cell layers of dorsal hippocampus that correlated with the behavioral CR. This increased neural activity, which the authors suggest may be one of the earliest neuronal indicators that learning has occurred, begins early in conditioning (in fact, during the first eight, CS–UCS pairings) and precedes the behavioral response by 35–40 msec. Initially, the neuronal response precedes the UCR, but as CRs begin to occur, the neuronal activity both increases and moves forward in the CS–UCS interval and always precedes the behavioral response by 35–40

msec. Furthermore, there was no increase in activity in unpaired controls, ruling out the possibility that the activity was due to pseudoconditioning, sensory registration, or motor output. A subsequent study (Berger & Thompson, 1978a) investigated single-unit activity in hippocampal pyramidal cells. As in the case of multiunit activity, the behavior of these cells was well correlated with the conditioned NMR.

Additional evidence for the role of the hippocampus in classical conditioning of the rabbit's NMR comes from a recent study by Berry and Thompson (1978). In this experiment, the authors were able to predict the rate of acquisition of the NMR by examining the hippocampal electroencephalogram (EEG) prior to conditioning. They found that animals that displayed relatively high levels of activity in the high-frequency category (8–22 Hz) conditioned more slowly than animals that showed a higher proportion of activity in the low-frequency range (2–8 Hz). Berry and Thompson suggested that this may be an indication that "behavioral state," as indicated by the hippocampal EEG, is an important factor in learning. Interestingly, Moore, Goodell, and Solomon (1976) reported that the cholinergic blocker scopolamine, which disrupts hippocampal activity in the low-frequency range (see Stumpf, 1965), greatly retards conditioning of the NMR to a tone or a light CS.

At first glance, the electrophysiological data on acquisition of the NMR may seem inconsistent with the results of lesion studies: If a structure is involved in a behavior, should not removal of the structure affect that behavior? Not necessarily. The finding that removal of a structure does not disrupt conditioning should not be interpreted to mean that the structure does not participate in the behavior in the intact animal. Furthermore, as Thompson (1976) pointed out in regard to the electrophysiological data, there is the possibility that brain structures besides the hippocampus (e.g., brainstem) may show similar changes in activity early in conditioning. Thus it may be that the hippocampus is monitoring the activity of other areas. If this is the case, hippocampal lesions would not necessarily disrupt acquisition of the CR. Finally, as mentioned earlier, the effects of hippocampal lesions have only been investigated in terms of crude dependent measures such as the presence or absence of a CR. A more fine-grained analysis of the effects of hippocampal ablation on conditioning of the NMR may yield differences.

Electrical Stimulation of the Brain

Like the data from lesion and electrophysiological studies, research using electrical stimulation of the brain also implicates the hippocampus in conditioning of the rabbit's NMR.

A recent study by Salafia, Romano, Tynan, and Host (1977) found that posttrial stimulation of the hippocampus immediately following the UCS offset retarded conditioning of the NMR. Once CRs began to occur, however, posttrial stimulation was without effect. The level of electrical stimulation in these studies did not produce convulsions, but it was sufficient to evoke poststimulation seizure activity in hippocampus. Based on these data, Salafia et al. (1977) concluded that the disruptive effects of posttrial hippocampal stimulation primarily affected association processes, as opposed to either the registration of the stimulus or the execution of the response. Subsequent work by Salafia (Salafia, Chiaia, & Ramirez, in press) indicates that the development of the CR can be as severely retarded if subseizure stimulation is used in either the dorsal hippocampus or amygdala.

Spatial and Temporal Information Processing in the Hippocampus?

Black et al. (1977) propose that the hippocampus is involved in the formation of spatial maps of the environment. The data and arguments presented by these authors as well as the data presented by researchers working in similar paradigms (e.g., Olton et al., 1978) suggest that this is indeed *one* function of the hippocampus. Nevertheless, the evidence from the rabbit NMR preparation indicates that at least in the rabbit, the hippocampus is involved in other processes.

Since most of the data that support the spatial information processing theory are based on the rat, it is tempting to speculate that the apparent discrepancies between these argu-

ments and those presented in this article are entirely due to species differences. Although differences in species and perhaps more importantly, differences in how dependent each species is on spatial information, are likely to account for some of the variability, it is equally likely that these differences do not account for the entire discrepancy. For example, Black et al. (1977, p. 1113) cite many studies that show facilitated avoidance learning in hippocampectomized animals. To explain this facilitation, they argue that these animals are not subject to interference from spatial cues. This facilitated avoidance, however, occurs not only in rats but also in hippocampectomized rabbits (Papsdorf & Woodruff, 1970). Similarly, we (Solomon & Moore, 1975) claimed, on the basis of rabbit NMR data, that the disruption of LI and blocking in animals with hippocampal lesions was due to the animal's inability to learn to ignore an irrelevant stimulus. These findings are not unique to the rabbit, as similar results have been reported for the rat in LI, using both the two-way avoidance (Ackil, Melgren, Halgren, & Frommer, 1969) and taste aversion paradigms (McFarland, Kostas, & Drew, 1978) and in blocking in the CER paradigm (Rickert, Bennett, Lane, & French, 1978). Furthermore, Patterson, Berger, and Thompson (1979) found that hippocampal-unit activity in the cat during NMR conditioning is similar to that found in the rabbit. This suggests that the electrophysiological evidence in the rabbit NMR preparation is also not species dependent. Thus, although it is possible that the rat is more of a spatial animal than the rabbit, especially the restrained rabbit, this potential difference cannot explain the apparent discrepancies in hippocampal function between the two species. (But see Winson, 1972, for a discussion of species differences and hippocampal function.) Rather, it seems that although the hippocampus may be involved in spatial information processing, it may be implicated as well in the processing of temporal information about stimuli.

Without attempting to add yet another all-encompassing theory of hippocampal function, I would like to suggest that in addition to the formation of spatial maps (and no doubt participation in a variety of other functions; see Isaacson & Pribram, 1975, pp. 429–442; Nadel & O'Keefe, 1974, for a review of the reviews) *one* function that the hippocampus might participate in is a type of "temporal mapping."[2] By this I simply mean registration of the temporal sequence of events. In the rabbit NMR preparation, this would be a coding of the relationship between the CS or the CSs and the UCS. Thus in LI the hippocampus receives information that a CS has occurred and that it is not followed by a relevant (i.e., motivationally significant) event, namely, the UCS. Consequently, the hippocampus participates in a process whereby the irrelevant stimulus is tuned out. One interesting way to test this idea of hippocampal function in LI would be to record from the hippocampus not only during stimulus preexposure but perhaps more importantly, during conditioning following the stimulus preexposure. Although this has yet to be done in the rabbit, Best and Best (1976) presented data from an LI paradigm in the rat that showed increased firing in hippocampal cells during conditioning for nonpreexposed animals but decreased activity in animals that had been preexposed.

In blocking, the animal learns a slightly more complex temporal sequence: CS_A always predicts the UCS, and any additional information (e.g., CS_B) in regard to the UCS is redundant and, like a preexposed stimulus, should be tuned out.

The notion of temporal information processing may also be consistent with the data on excitatory conditioning despite the finding that hippocampal lesions do not disrupt this form of learning. In excitatory conditioning the hippocampus may receive information that a CS has occurred and that it is closely followed by a relevant event, the UCS. This temporal sequence is coded in the hippocampus and this coding may account for the change in neural activity reported by Thompson and his coworkers during CS–UCS pairings. In fact, Thompson notes (cf. Berger & Thompson, 1978b) that hippocampal-unit activity actually

[2] I am grateful to John W. Moore of the University of Massachusetts at Amherst for discussions on temporal information processing in the hippocampus. See Moore (1979), who proposes a neural model to account for many of the phenomena discussed in this article.

forms a temporal model of the temporal form of the behavioral NMR. Since the CS is paired with a relevant event, the UCS, the hippocampus should not act to tune it out. Furthermore, although a hippocampectomized animal loses the ability to tune out irrelevant stimuli, this would have no effect on simple acquisition, since tuning out an irrelevant stimulus is not required. Thus, whereas the hippocampus receives information about the sequence of events transpiring during excitatory conditioning, it need not initiate action, since the CS is relevant. Consequently, hippocampal lesions have no effect on acquisition of the CR. This is not to say that the hippocampus is not critical to acquisition of the CR under certain circumstances. For example, animals with hippocampal lesions should be more susceptible to the deleterious effects of external inhibitors (distractors) on conditioning, and there is some preliminary data to suggest that this is the case (Solomon & Moore, 1975, Experiment 2).

Salafia et al.'s (1977) posttrial stimulation data also fit nicely into this scheme. Hippocampal stimulation could result in the tuning out of all stimuli, both relevant and irrelevant, and would thus have the effect of retarding conditioning.

In summary, it may be that the hippocampus is capable of establishing the relationship (or to use Black et al.'s terminology, "maps" the relationship) between many kinds of stimuli. Certainly the seemingly redundant lamellar organization of the structure (e.g., Anderson, Bliss, & Skrede, 1971) coupled with its anatomical relationships to sensory and "motivational" systems of the brain suggests this potential. What type of mapping occurs, however, may depend on what animal is being investigated, what type of stimuli are available, and what type of problem the animal is asked to solve.

Reference Notes

1. Patterson, M. M. Personal communication, January 1978.
2. Moore, J. W., Mis, F. W., & Solomon, P. R. *Hippocampal and midbrain mechanisms in inhibition of the rabbit's nictitating membrane response.* Paper presented at the meeting of the Psychonomic Society, Denver, Colo., November 1975.

References

Ackil, J. R., Melgren, R. L., Halgren, C., & Frommer, S. P. Effects of CS preexposure on avoidance learning in rats with hippocampal lesions. *Journal of Comparative and Physiological Psychology*, 1969, *69*, 739–747.

Anderson, P., Bliss, T. V. P., & Skrede, K. K. Lamellar organization of hippocampal excitatory pathways. *Experimental Brain Research*, 1971, *13*, 222–238.

Berger, T. W., Alger, B. E., & Thompson, R. F. Neuronal substrates of classical conditioning in the hippocampus. *Science*, 1976, *192*, 483–485.

Berger, T. W., & Thompson, R. F. Identification of pyramidal cells as the critical elements in hippocampal neuronal plasticity during learning. *Proceedings of the National Academy of Science*, 1978, *75*, 1572–1576. (a)

Berger, T. W., & Thompson, R. F. Neuronal plasticity in the limbic system during classical conditioning of the rabbit's nictitating membrane response: 1. The hippocampus. *Brain Research*, 1978, *145*, 323–336. (b)

Berry, S. D., & Thompson, R. F. Prediction of learning rate from hippocampal EEG. *Science*, 1978, *200*, 1298–1300.

Best, M. R., & Best, P. J. The effects of state of consciousness on latent inhibition in hippocampal unit activity in the rat during conditioning. *Experimental Neurology*, 1976, *51*, 564–573.

Black, A. H., Nadel, L., & O'Keefe, J. Hippocampal function in avoidance learning and punishment. *Psychological Bulletin*, 1977, *84*, 1107–1129.

Douglas, R. J. The hippocampus and behavior. *Psychological Bulletin*, 1967, *67*, 416–442.

Frey, P. W., Maisiak, R., & Dugue, G. Unconditional stimulus characteristics in rabbit eyelid conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, 1976, *2*, 175–190.

Gormezano, I. Classical conditioning. In J. B. Sidowski (Ed.), *Experimental methods and instrumentation in psychology*. New York: McGraw-Hill, 1966.

Hearst, E. Some persistent problems in the analysis of conditioned inhibition. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and learning*. New York: Academic Press, 1972.

Isaacson, R. L., & Pribram, K. H. *The hippocampus* (Vol. 2). New York: Plenum Press, 1975.

Kamin, L. J. "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior*. Miami: University of Miami Press, 1968.

Kamin, L. J. Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior*. New York: Appleton-Century-Crofts, 1969.

Kimble, D. P. Hippocampus and internal inhibition. *Psychological Bulletin*, 1968, *76*, 285–295.

Mackintosh, N. J. Stimulus selection: Learning to ignore stimuli that predict no change in reinforcement. In R. A. Hinde & J. Stevenson-Hinde (Eds.), *Constraints on learning*. New York: Academic Press, 1973.

Mackintosh, N. J. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 1975, *82*, 276–298.
Marchant, H. G., Mis, F. W., & Moore, J. W. Conditioned inhibition of the rabbit's nictitating membrane response. *Journal of Experimental Psychology*, 1972, *95*, 408–411.
Marchant, H. G., & Moore, J. W. Blocking on the rabbit's conditioned nictitating membrane response in Kamin's two-stage paradigm. *Journal of Experimental Psychology*, 1973, *101*, 155–158.
Marchant, H. G., & Moore, J. W. Below-zero conditioned inhibition of the rabbit's nictitating membrane response. *Journal of Experimental Psychology*, 1974, *102*, 350–357.
McFarland, D. J., Kostas, J., & Drew, W. G. Dorsal hippocampal lesions: Effects of preconditioning CS exposure on flavor aversion. *Behavioral Biology*, 1978, *22*, 398–404.
Mis, F. L. A midbrain-brain stem circuit for conditioned inhibition of the rabbit's (*Oryctolagus cuniculus*) nictitating membrane response. *Journal of Comparative and Physiological Psychology*, 1977, *91*, 975–980.
Moore, J. W. Brain processes and conditioning. In A. Dickinson & R. A. Boakes (Eds.), *Associative mechanisms in conditioning*. Hillsdale, N.J.: Erlbaum, 1979.
Moore, J. W., Goodell, N. A., & Solomon, P. R. Central cholinergic blockade by scopolamine and habituation, classical conditioning, and latent inhibition of the rabbit's nictitating membrane response. *Physiological Psychology*, 1976, *4*, 395–399.
Nadel, L., & O'Keefe, J. The hippocampus in pieces and patches: An essay on modes of explanation in physiological psychology. In R. Bellairs & E. G. Gray (Eds.), *Essays on the nervous system: A festschrift for Professor J. Z. Young*. Oxford, England: Clarendon Press, 1974.
Nadel, L., O'Keefe, J., & Black, A. H. Slam on the brakes: A critique of Altman, Brunner, and Bayer's response-inhibition model of hippocampal function. *Behavioral Biology*, 1975, *14*, 151–162.
O'Keefe, J. Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, 1976, *51*, 78–109.
O'Keefe, J., & Black, A. H. Single unit and lesion experiments on the sensory inputs to the hippocampal cognitive map. In, *Functions of the septo-hippocampal system, CIBA Foundation Symposium 58*. New York: American Elsevier, 1979.
O'Keefe, J., & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. *Brain Research*, 1971, *34*, 171–175.
O'Keefe, J., Nadel, L., Kieghtly, S., & Kill, D. Fornix lesions selectively abolish place learning in the rat. *Experimental Neurology*, 1975, *48*, 152–166.
Olton, D. S., Branch, M., & Best, P. J. Spatial correlates of hippocampal unit activity. *Experimental Neurology*, 1978, *58*, 387–409.
Olton, D. S., Walker, J. A., & Gage, F. H. Hippocampal connections and spatial discrimination. *Brain Research*, 1978, *139*, 295–308.
Papsdorf, J. D., & Woodruff, M. Effects of bilateral hippocampectomy on the rabbit's acquisition of shuttle box and passive avoidance responses. *Journal of Comparative and Physiological Psychology*, 1970, *73*, 486–489.
Patterson, M. M., Berger, T. W., & Thompson, R. F. Neural plasticity recorded from cat hippocampus in classical conditioning. *Brain Research*, 1979, *163*, 339–343.
Rescorla, R. A. Pavlovian conditioned inhibition. *Psychological Bulletin*, 1969, *72*, 77–94.
Rickert, E. J., Bennett, T. L., Lane, P., & French, J. Hippocampectomy and attenuation of blocking. *Behavioral Biology*, 1978. *22*, 147–160.
Salafia, W. R., Chiaia, N. L., & Ramirez, J. J. Retardation of rabbit nictitating membrane response conditioning by subseizure electrical stimulation of hippocampus. *Physiology and Behavior*, in press.
Salafia, W. R., Romano, A. G., Tynan, T. T., & Host, K. C. Disruption of rabbit (*Oryctolagus cuniculus*) nictitating membrane conditioning by post-trial electrical stimulation of hippocampus. *Physiology and Behavior*, 1977, *18*, 207–212.
Schmaltz, L. W., & Theios, J. Acquisition and extinction of a classically conditioned response in hippocampectomized rabbits (*Oryctolagus cuniculus*). *Journal of Comparative and Physiological Psychology*, 1972, *79*, 328–333.
Solomon, P. R. Role of the hippocampus in blocking and conditioned inhibition of the rabbit's nictitating membrane response. *Journal of Comparative and Physiological Psychology*, 1977, *91*, 407–417.
Solomon, P. R., & Moore, J. W. Latent inhibition and stimulus generalization of the classically conditioned nictitating membrane response in rabbits. (*Oryctolagus cuniculus*) following dorsal hippocampal ablation. *Journal of Comparative and Physiological Psychology*, 1975, *89*, 1192–1203.
Stumpf, C. The fast component in the electrical activity of the rabbit's hippocampus. *Electroencephalography and Clinical Neurophysiology*, 1965, *18*, 477–486.
Thompson, R. F. The search for the engram. *American Psychologist*, 1976, *31*, 209–227.
Wagner, A. R., Rudy, J. W., & Whitlow, J. E. Rehearsal in animal conditioning. *Journal of Experimental Psychology*, 1973, *97*, 407–426.
Winson, J. Interspecies differences in the occurrence of theta. *Behavioral Biology*, 1972, *7*, 479–487.

Received June 26, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1280–1289

Tuning Out Irrelevancy? Comments on Solomon's Temporal Mapping View of the Hippocampus

John O'Keefe
Department of Anatomy
University College London, London, England

Lynn Nadel and Jeff Willner
Dalhousie University, Nova Scotia, Canada

In his article, Solomon proposes an extension to the cognitive map theory of hippocampal function. He cites evidence drawn mainly from the literature on the classical conditioning of the rabbit's nictitating membrane response (NMR), which he claims can not be satisfactorily explained within the framework of the present theory. Although rabbits with hippocampal lesions acquire the conditioned response as rapidly as do normals and show normal conditioned inhibition, they do not show latent inhibition or blocking. Single-unit and multiple-unit recordings during NMR conditioning show that some hippocampal cells begin to increase their firing rates during the conditioned stimulus period as learning takes place. Posttrial electrical stimulation of the hippocampus retards NMR conditioning. To explain these findings, Solomon proposes that in addition to its spatial mapping function, the hippocampus might also function as a temporal map. A temporal map would register the temporal relationship between stimuli and "tune out" irrelevant stimuli. In this article we examine the studies that Solomon cites as incompatible with the spatial map theory. We show that the deficits following hippocampal lesions can be predicted from an understanding of the role of place learning in each paradigm, and we show that the electrophysiological data are open to different interpretation. We conclude that the data cited by Solomon do not require any modification of the theory.

In his article, Solomon (1979) comments on our spatial map theory of hippocampal function and points to what he sees as its inadequacies in dealing with a certain set of data. Although he allows that "the spatial information processing view of hippocampal function accounts for much of the available data" (p. 1272), he argues that "it cannot account for the data from the classically conditioned rabbit nictitating membrane response preparation" (p. 1272). After reviewing data from both lesion and electrophysiological experiments on the rabbit's nictitating membrane response (NMR), he concludes that our theory needs to be modified: In addition to its spatial mapping function, the hippocampus must also participate in a type of temporal mapping. During NMR conditioning, a temporal map would encode the temporal relationship between the conditioned stimulus (CS) or CSs and unconditioned stimulus (US) and participate in a process whereby irrelevant stimuli (ones not signaling changes in reinforcement contingencies) are ignored or "tuned out." He thus seeks to broaden the information content of the map and at the same time re-

---

The research of J. O'Keefe was supported by the Medical Research Council of England, and the research of L. Nadel was supported by the Natural Sciences and Engineering Research Council of Canada.

The authors would like to thank D. Conway, who collaborated on the research described herein. This reply to a critique of an article written with Abe Black is respectfully dedicated to his memory.

Requests for reprints should be sent to J. O'Keefe, Cerebral Functions Group, Department of Anatomy, University College London, Gower Street, London WC1E 6BT, England.

strict its output function to the inhibitory one of filtering out sensory stimuli.[1]

Although we are not against modification and extension of the theory in principle (see, for example, our discussion of semantic maps in O'Keefe & Nadel, 1978), we feel compelled to resist Solomon's particular modification. The reasons for doing so are set out in an earlier article on the distinction between theories and hypotheses (Nadel & O'Keefe, 1974). In brief, the power of the theory as it now stands derives from its ability to make clear predictions that can be tested and judged to be correct or incorrect. Solomon's modification would overextend the theory's explanatory power at the expense of its predictive power. Let Solomon say it: "In summary, it may be that the hippocampus is capable of establishing the relationship (or to use Black et al.'s terminology, "maps" the relationship) between many kinds of stimuli . . . . What type of mapping occurs, however, may depend on what animal is being investigated, what type of stimuli are available, and what type of problem the animal is asked to solve" (pp. 1278). In short, the new hybrid theory has been reduced to an hypothesis, all of the important questions to be answered by ad hoc postulates. Before allowing that to happen, we must examine carefully Solomon's assertions and the data on which they rest. We try to show that these data are only apparently in conflict with the cognitive map predictions. His article is divided into four parts and we follow that order in our comments. In the first part (the section on The Spatial Information Processing Hypothesis and Classical Conditioning and the section on Absence of Spatial Cues in the Rabbit NMR Preparation), Solomon discusses our spatial map theory and concludes that there is little or no role for the mapping system in classical conditioning of the rabbit NMR. We are at pains to deny this and point to several ways in which the map could be involved and how these ideas can be tested. In the second part of his article (the section on Lesion Studies), Solomon reviews experiments on the effects of hippocampal lesions on classical conditioning, especially that of the NMR. He concludes that lesions do not affect the acquisition of simple excitatory conditioning or the development of conditioned inhibition but that they do disrupt latent inhibition and blocking. Here we look more closely at the role of background cues (i.e., places) in these paradigms and conclude that the data do not support Solomon's interpretations. In his third part (the sections on Electrophysiological Studies and Electrical Stimulation of the Brain), Solomon cites recent experiments that show changes in multiple-unit and single-unit recordings from the hippocampus during classical conditioning of the NMR. We have commented briefly on these studies elsewhere (O'Keefe & Nadel, 1978, pp. 190–194) and extend those comments here. In a final part (the section on Spatial and Temporal Information Processing in the Hippocamus), Solomon gives a resume of his own notion of a temporal map. We feel that the notion as presented has little explanatory power and constitutes an unjustified and unnecessary extension of the hippocampal cognitive map theory.

Place Cues: Their Role in Classical Conditioning

Solomon asserts that place learning plays little if any role in classical conditioning of the rabbit NMR because "the animal remains virtually motionless throughout the conditioning session, and the conditioned stimuli (CS) and the unconditioned stimuli (UCS) are delivered in the same spatial location at all times" (p. 1273). Solomon may be forgiven for misunderstanding our position, since we have not emphasized the role of place learning in classical conditioning, but we are somewhat surprised that he has chosen to ignore the recent experimental and theoretical literature that points to the importance of background cure or places in classical conditioning.[2]

---

[1] It is not clear whether Solomon intends the restriction of the output function to be applied only to the enlarged map's temporal functions or to the spatial functions as well. We have presented the arguments against a general inhibitory function for the hippocampus elsewhere (e.g., Nadel & O'Keefe, 1974; Nadel, O'Keefe, & Black, 1975) and will not repeat them here.

[2] One gets the impression that the term "background cues" is used to refer to a collection of individual cues that, although unidentified, do not differ in principle from the specific foreground CSs. We have argued that this view is only partially correct: In most experimental situations the major role of the background

According to the cognitive map theory, there are three different ways that the hippocampus can be involved in learning and performance.

1. Learning about places: Organisms built maps of environments. An environment or specific places within it can be tagged as dangerous or as containing rewards such as food or water.

2. Strategies using places: Using its map, the animal can identify its position within the environment as well as the distance and direction to other places.

3. Attention to places: The mapping system can aid in learning about a stimulus in an environment either by directing the animal's attention to particular places in which that stimulus occurs or by assessing the stimulus as novel, that is, one that is not represented in the mapping system as occurring in that place in the environment.

One way in which Pavlovian tasks differ from instrumental tasks is that they rule out the second use of the mapping system as a successful strategy for solving problems; by definition, nothing the animal does can affect the contingency between the CS and US. This does not mean that the animal won't try to use such place strategies; under certain circumstances they may actually conflict with what the experimenter is measuring as learning.

Several experiments have pointed to the role of place factors in Pavlovian conditioning. Sheffield (1965) reported that during classical conditioning of the salivary response, dogs would often emit bursts of salivation during the intertrial interval:

> These bursts are such an ever-present nuisance that the experimenter, who is watching the dog, the polygraph, and the tape that programs trials, is kept in a continual state of anxiety lest a burst start just ahead of CS, spoiling the record of what otherwise might be a clear cut CR. Also it is often difficult to tell whether a "CR" was actually a response to the CS or whether it was a lucky burst, accidentally timed with the CS. Moreover, the bursts between trials look so much like the response to CS that one is obliged to raise the question of whether the bursts also have a "CS" and if so, what the nature of this unobserved CS is. (p. 314)

These bursts were maximal early in conditioning, declining somewhat later in training. Sheafor (1975) has studied such pseudoconditioned responses during classical conditioning of the rabbit's jaw movement using a water US. He concluded that they resulted from an association of the background cues with the US and were not due to trace or temporal conditioning to the specific foreground CS. Part of his evidence was that these pseudoconditioned CRs could be extinguished by leaving the animal in the test situation without presentation of the CS or US.

The most straightforward way to test for the development of a place hypothesis during classical conditioning is to carry out probe trials during which the animal is tested in a different environment or is freed from restraint and put either in the usual testing place or in a new nearby place. Zener (1937) was the first to do these latter tests and some of his results might be interpreted as support for place learning. When put unrestrained on a new testing stand and presented with a CS signaling food, the dog left the new stand, went over to the usual stand and waited for the food to be delivered. There is some indication that the animal would have moved from the new table to the old in the absence of the CS but was prevented from doing so by Zener. In response to a different CS that signaled acid in the mouth, the same animal first left the new platform and returned to the usual testing place but then turned around and walked away from it.

In certain classical conditioning paradigms, this kind of place learning seems to conflict with conditioning to the experimenter-designated CS. Rescorla and Wagner (1972) have suggested that a particular US has a limited amount of associative strength and that this must be shared amongst all of the stimuli conditioned to it. The amount of conditioning to each stimulus depends on how good it is at predicting the US in relation to other stimuli in the situation. In the standard classical conditioning paradigm, this associative strength is, at least initially, shared between the background and a compound consisting of the

---

cues is to identify environments or specific parts of an environment. These *places* have some properties in common with foreground CSs but differ in many important respects (see O'Keefe & Nadel, 1978, pp. 80–101). Here we assume that most of the effects attributed to background cues are due to their role in defining places.

background cues and the specific foreground CS. Manipulations that alter the strength of conditioning to the background cues should have an inverse effect on conditioning to the specific foreground CS and vice versa. Whether one accepts the notion that there is a limited amount of associative strength to be shared (see Mackintosh, 1975, for an alternative view), it seems clear that spatial context can influence conditioning to specific foreground CSs. Several experiments provide support for this idea.

In a conditioned emotional response (CER) study, Dweck and Wagner (1970) showed that extinguishing the fear associated with the background cues by exposing the animal to the environment in the absence of the CS and US during training increased conditioned suppression to the specific CS. Odling-Smee (1978) confirmed this finding in a somewhat different situation, whereas Odling-Smee (1975) showed the converse: Increasing the correlation between the specific CS and US reduced the aversiveness of the place in which conditioning occurred, even though establishing a perfect correlation between CS and US did not totally eliminate fear of the training environment. Taken together, these studies support the notion that fear can be conditioned to the environment and suggest that the degree of such conditioning is influenced by the reliability with which the specific stimuli and the spatial context (i.e., background cues) predict the US.

In addition to participating in conditioning in the manner previously described, the mapping system also can become involved through its misplace subsystem. This component detects discrepancies between the stored representations of the spatial array of stimuli that occur at a particular place in an environment and those that the animal perceives at any given moment. Mismatches generate exploration that serves to incorporate the changes into the new representation of that place. We have argued that the introduction of a new stimulus into an environment has two kinds of effects: One effect is due to its novelty, which generates a mismatch in the hippocampus, and the other is due to its noticeability, which depends on the properties of the sensory analyzers, how intense the stimulus is, how recently it or a similar stimulus has been experienced, and so forth.

Unlike noticeability, the novelty of a stimulus depends on the environmental context in which it occurs. This being the case, the mapping system should be involved in those aspects of Pavlovian conditioning for which novelty is a critical factor. In latent inhibition (LI), for instance, experience with a stimulus and hence a reduction in its novelty retards its subsequent associability with a US (e.g., Lubow, 1973). This loss of associability seems limited to the context in which exposure to the stimulus occurred. Lubow, Rifkin, and Alek (1976) looked at the effect of exposure to olfactory stimuli in the same or different environments. LI was only obtained when the CS was preexposed in the same environment in which conditioning subsequently occurred. There was no retardation of conditioning to a CS that had been preexposed in a different environment. These data suggest that the hippocampal mapping system is involved in LI: By defining the context of occurrence, the map makes novelty possible.

In sum, the hippocampal cognitive map theory predicts that the hippocampus is involved in classical conditioning, often in rather subtle ways. We can now turn to the changes that Solomon cites as inexplicable within the framework of the theory as it stands.

No Change in NMR Acquisition After Hippocampal Lesions

Solomon concentrates on the effects of lesions on various conditioning paradigms involving the rabbit NMR but also makes reference to other conditioning studies. He admits that hippocampal lesions do not affect acquisition of the NMR, which is in line with the mapping theory, but suggests that more sensitive response measures might reveal subtle changes in amplitudes, latency, or interstimulus interval (ISI) shifts. In view of Sheafor's (1975) evidence that CRs can be elicited and influenced by background cues, we concur. It is not immediately apparent, however, how Solomon's ideas would explain such effects.

Hippocampal Lesions, LI, and Blocking

We now turn to a closer examination of the conditioning paradigms in which hippocampal lesions do produce deficits and to the interpretation Solomon places on these data. We begin with an analysis of LI, briefly discuss blocking, and then move on to conditioned inhibition, which provides a critical test of our respective positions. Our argument throughout will be that (a) place learning does have a role in certain Pavlovian procedures, and (b) the effects of a hippocampal lesion on an animal's performance in any given paradigm can be predicted from a knowledge of the role of place learning in that procedure.

Solomon and Moore (1975) have shown that LI in NMR conditioning is disrupted by hippocampal lesions. Solomon (1979) cites two other studies that have found a failure of LI in different paradigms as evidence of the generality of the effect (Ackil, Mellgren, Halgren, & Frommer, 1969, in two-way avoidance, and McFarland, Kostas, & Drew, 1978, in taste aversion learning[3]).

Our theory already predicts an effect of hippocampal lesions on LI insofar as it is dependent on environmental context; there is no need to extend the theory in the way suggested by Solomon. Moreover, taken as it stands, the cognitive map theory suggests a more comprehensive interpretation of this phenomenon. One LI study that Solomon failed to cite was reported by Olton and Isaacson (1968). Superficially similar to the Ackil et al. study cited by Solomon in that it used a two-way avoidance task, this study found that preexposure to the CS facilitated rather than retarded subsequent learning in intact rats. In neither study did preexposure influence learning in the lesioned rats. The failure to obtain preexposure effects in the lesioned animals is explicable; the opposed effects of CS preexposure in the intact rats in these two studies is less so.

Analysis of the two-way task in terms of the role of place and other kinds of learning helps to clarify this apparent discrepancy. The normal intact animal has considerable difficulty in learning two-way avoidance. In Black, Nadel, and O'Keefe (1977) we argued that this is due to the fact that place hypotheses hinder learning, since (a) they promote freezing in inescapable, dangerous places, and (b) they retard movement back into a dangerous compartment in which the animal has recently been shocked. In support of this view, it has been shown that manipulations that increase the discriminability of the two compartments of the shuttle box will exacerbate the intact animal's difficulties (see O'Keefe & Nadel, 1978, p. 301). The robust facilitation of learning the two-way task that occurs after hippocampal lesions are made offers added support; without the possibility of (maladaptive) place strategies, these rats acquire the task easily. In contrast to this debilitating effect of place learning, specific foreground CSs, especially if they act as directional cues that the animal can approach or avoid, will enhance learning.

The effects of preexposure on learning in the two-way task must be interpreted in terms of what is being preexposed and in terms of the role this component typically plays in learning. Preexposure to the place will reduce the likelihood of place strategies being used in any subsequent conditioning: By virtue of this preexposure the place becomes a relatively unreliable predictor of the US. Thus place preexposure should enhance solution of the two-way avoidance task. Preexposure to the specific cues will reduce their role in conditioning within that context because they will have been incorporated into the hippocampal representation of that environment, and there will be no misplace output to identify them as novel when they are reintroduced with the US. To put it another way, they will have been absorbed or embedded into the context and as a consequence, they are unlikely to be rapidly associated with the newly introduced US. Since such CSs normally enhance two-way avoidance

[3] McFarland et al. compared preexposed and nonpreexposed hippocampals and controls on their ability to form an aversion to a walnut solution paired with lithium carbonate. Their conclusion that hippocampal rats are deficient in LI was based on separate tests that were conducted with the walnut solution and tap water on the days immediately following poisoning. Inspection of their data reveals that the poisoning markedly depressed baseline water consumption of most of their rats. Any conclusions based on a comparison of water and walnut solution intakes are tenuous.

learning, their preexposure should act to retard it.

One major difference between the Ackil et al. (1969) and the Olton and Isaacson (1968) studies was in their control treatments for preexposure. In both studies, the experimental group was preexposed to background cues and specific foreground cues. However, this preexposed group was compared to different control groups in the two studies. Olton and Isaacson used a no-exposure control, whereas Ackil et al. used a place exposure control. This can be schematized as follows:

	Preexposed components		
	Place	Place + CS	Nothing
Olton & Isaacson	—	Experimental >	Control
Ackil et al.	Control >	Experimental	—

This layout shows the conditions used in the two studies, and the signs indicate the results. Preexposure to the place facilitates learning; preexposure to both place and CS facilitates learning somewhat less. This pattern of results makes sense within cognitive map theory, as does the absence of any preexposure effects in rats with hippocampal lesions.

Another paradigm discussed by Solomon is blocking. The basic procedure here involves first conditioning to one CS ($CS_A \rightarrow US$) and then presenting an added CS in compound with the pretrained CS during subsequent trials ($CS_ACS_B \rightarrow US$). The typical finding is that conditioning to CS_B is weaker than would occur in the absence of prior conditioning to CS_A. It is often said that because CS_B fails to predict anything not already predicted by CS_A, it is tuned out, and little conditioning occurs. There are defects in blocking in hippocampally lesioned subjects, as Solomon notes (Solomon, 1977; Rickert, Bennett, Lane, & French, 1978). There is as yet no fully accepted explanation of blocking.[4] In particular, the role of context in the development of blocking has yet to be determined. The present analysis predicts that it should have an important role, similar to that seen in LI. We treat this specific issue and provide further discussion of several Pavlovian paradigms, in another article (Nadel & Willner, Note 2).

Hippocampal Lesions and Conditioned Inhibition

A conditioned inhibitor (CS^-) is defined as a stimulus that as a result of certain prior experiences (specified later) (a) detracts from the efficacy of a CS known to elicit a conditioned response, and (b) itself conditions more slowly than in the absence of this prior experience (Rescorla, 1969). Rescorla and Wagner (1972; Wagner & Rescorla, 1972) have specified the conditions necessary to produce a CS^-: A CS will become a CS^- when it is negatively correlated with the US. That is, a stimulus will become an inhibitor when it signals a lower probability of US occurrence than the animal would otherwise expect.

According to Solomon (1979), it is precisely the fact that the CS^- signals a change in the conditions of reinforcement that prevents it from being tuned out, and this in turn prevents hippocampal lesions from having any effect on the establishment of conditioned inhibition. Solomon (1977) tested this notion in rabbits and found no differences between lesioned and control subjects. So far, these data appear to support Solomon's assertions. He also assumed that the spatial mapping theory predicts no general defects in this task. We agree with Solomon insofar as we feel that the hippocampus has no general inhibitory function. However, this does not mean that the hippocampus has no role in the learning process that underlies conditioned inhibition. To the extent that (spatial) context is important in the conditioning of inhibition, we expect defects in this paradigm in lesioned subjects. There are in fact a variety of procedures that will turn a CS into a CS^-, and it is useful to look at these procedures with

[4] Failures of blocking ("unblocking") can occur when the added CS is correlated with some change in stimulus conditions. For example, changes in US intensity (Kamin, 1969), or in the number of USs (Dickinson, Hall, & Mackintosh, 1976) produce unblocking. Brief, nonreinforced presentation of the CS_ACS_B compound also produces unblocking (Gray & Appignanesi, 1973). Interestingly, changes in the temporal duration (Kohler & Ayres, Note 1) or temporal patterning (Gray, 1978) of the CSs fail to produce unblocking. This set of results is not particularly congenial to a temporal mapping/tuning out interpretation of the hippocampal role in blocking.

an eye towards the role that place learning plays in each of them.

Solomon (1977) generated conditioned inhibition through the use of the classical Pavlovian procedure (Pavlov, 1927). In this procedure the animal receives trials in which CS_A is always followed by the US, interspersed with trials in which a compound of CS_A and CS_B is never followed by the US. As a result of this training, the added CS_B acquires substantial inhibitory properties (i.e., becomes a CS^-). In this preparation, an expectancy of the US is generated by the presentation of CS_A (the CS^+). When the compound of CS_A and CS_B is not followed by the US, this expectancy is not confirmed. The added CS becomes an inhibitor because it is correlated with the omission of an otherwise expected US. Background cues play little role in this case and hippocampal lesions accordingly have little effect.

However, there are other paradigms generating conditioned inhibition in which spatial context seems to be of greater importance, such as those in which the CS and US are explicitly unpaired or in so-called discriminative classical conditioning. In their analysis of conditioned inhibition, Wagner and Rescorla (1972) argued that the development of inhibition in these preparations is mediated by the context (place) in which conditioning occurs. That is, the expectancy of the US is generated by the place. Since the hippocampus is crucial to learning about places, we predict deficits in lesioned subjects when these paradigms are used. It is of some interest that the other published study on conditioned inhibition in hippocampal animals employed the discriminative conditioning procedure. Micco and Schwartz (1971) trained rats in a situation in which one CS (CS^+) was always followed by a shock US, whereas a second CS (CS^-) was never followed by the shock CS. They then tested these CSs on an avoidance baseline. In normal rats the CS^+ facilitated avoidance responding (presumably through enhanced fear), whereas the CS^- suppressed responding. Hippocampal rats, on the other hand, showed increased responding in the presence of the CS^+ but virtually no effects of CS^- presentation. In other words, when places are important in the development of conditioned inhibition, hippocampal lesions produce profound defects. It is important to note that tuning out plays no role in any of these conditioned inhibition paradigms, according to Solomon.[5]

Thus, the available data support Solomon's contention that the hippocampus is involved in certain Pavlovian procedures but fail to support his assertion that this involvement demands an enlarged view of hippocampal function that goes beyond space into time. Except for the untested case of blocking, spatial context has been implicated as an important factor in those conditioning paradigms in which hippocampal lesion effects have been demonstrated. Furthermore, a defect was seen in a version of conditioned inhibition training in which place learning is important but in which the tuning out of irrelevancy is not. In sum, then, the lesion data discussed by Solomon offer no strong reason to abandon the parsimonious view that the hippocampus is merely a spatial mapping system. Rather, the lesion data argue for a closer examination of the role of spatial context in tasks that have hitherto been thought of as relatively free of such influences.

Hippocampal Recording During NMR

A second line of evidence that Solomon cites to support his notion that the hippocampus is involved in NMR is the changes in hippocampal unit firing during NMR reported by Berger, Thompson, and their colleagues (Berger, Alger, & Thompson, 1976; Berger & Thompson, 1978a, 1978b; Thompson, 1976).

We have worried about the apparent discrepancy between the results of single-unit studies in the hippocampus of rat and rabbit. Most recent studies on the rat (e.g., Best & Ranck, 1975; Hill, 1978; O'Keefe, 1976; O'Keefe & Conway, 1978; O'Keefe & Dostrovsky, 1971; Olton, Branch, & Best, 1978) report that many hippocampal cells fire preferentially to one part of an environment (*place cells*). In contrast, studies on the rabbit do not report place cells but concentrate on the changes in unit firing during classical conditioning

---

[5] This analysis of the Micco and Schwartz study makes sense of what previously seemed to us an anomalous result (O'Keefe & Nadel, 1978).

(Thompson's group) or on the response of hippocampal units to simple sensory stimuli such as flashing lights and pure tones (e.g., Vinogradova, 1970, 1975, see also the discussion in Elliott & Whelan, 1978, pp. 173–175, 192, 197). There might be several reasons for this discrepancy. One is that the hippocampus has a different function in different species. Winson (1972) has produced a variation of this argument with respect to hippocampal theta. This appears to be the most radical solution, and we agree with Solomon that it is unlikely to be the case. More probably, the investigators of rabbit and rat are looking at different aspects of the cell responses or are looking at the same cells but interpreting their responses differently. Two concrete possibilities immediately spring to mind. One possibility is that the groups working on the rabbit record cells while the animal is restrained in a small featureless box. If there are place cells in the rabbit hippocampus, this is the worst possible environment in which to find them, and even if they did have fields in these environments, the testing procedures would not allow the experimenters to see this. A second possibility is that the theta cells in the rabbit might change their firing rates during arousal as well as during movement. Support for this idea can be drawn from the observation that theta occurs in the rabbit in response to sensory stimuli as well as during movement (e.g., Harper, 1971; Kramis, Vanderwolf & Bland, 1975).

To check these possibilities, O'Keefe (Note 3) has recorded units from fields CA1 and FD[6] of the hippocampus of the freely moving rabbit. Preliminary results show that there are both place cells and theta cells in the rabbit hippocampus. The rabbit place cells seem identical to those recorded in the rat, whereas the theta cells in the rabbit fire during arousal as well as during movement, in contrast to the rat's theta cells. Berger and Thompson (1978a) have reported on some of the physiological characteristics of the hippocampal cells that were involved in conditioning. Most of the units that were involved could be antidromically activated by electrical shocks to the fornix and (judging from the histograms presented) appeared to have high spontaneous rates. On the other hand, units not involved in conditioning could not be antidromically activated from the fornix; some were orthodromically activated while the rest were not affected at all. This latter group "tended to have very low spontaneous rates, sometimes showing interspike intervals as long as 60–120 sec" (p. 1574). The spontaneous rates suggest that this last group is composed of place cells, whereas the high firing cells involved in conditioning are primarily theta units. However, the finding that some of the units involved in conditioning can be antidromically activated from the fornix and are therefore probably pyramidal cells is not consistent with this analysis. There is evidence in the rat that the complex spike cells (into which category the place cells fall) and not the theta cells are the pyramidal cells (Fox & Ranck, 1975, 1977).

Hippocampal Stimulation During NMR

Finally, Solomon cites the stimulation studies of Salafia and his colleagues (Salafia, Chiaia, & Ramirez, 1979; Salafia, Romano, Tynan, & Host, 1977) as support for his tuning out hypothesis. Salafia found that stimulation of the hippocampus (and more recently the amygdala) after each CS–US pairing retarded conditioning of the NMR. Solmon thinks that stimulation could result in the tuning out of all stimuli, relevant or irrelevant, and in this way retard conditioning. How sound is this reasoning? Since hippocampal lesions have no effect on acquisition, Solomon must think that the stimulation is not producing a temporary lesion but is actively blocking sensory transmission. But in Salafia's studies the stimulation is given after both the CS and US have been presented and cannot have any role in tuning out either on that trial. The experiment that *would* support Solomon's tuning out hypothesis would be one in which hippocampal stimulation during or just prior to the CS retarded conditioning, whereas stimulation at other times had no effect. Stimulation after the trial would be a control condition designed to rule out general stimulation effects, consolidation effects, prograde effects, and so forth. In general, we suggest caution in the

[6] CA1 = cornu ammonis 1, FD = fascia dentata.

interpretation of stimulation effects, since they can affect wide areas of the brain and when afterdischarges are involved, can cause long lasting changes in distant synapses (Goddard, McIntyre, & Leech, 1969; Racine, 1972).

In summary, Solomon has called attention to a set of data on classical conditioning that we have not previously dealt with adequately and that he felt lay outside the explanatory domain of the hippocampal cognitive map theory. He suggested that the theory is inadequate and needs to be extended to include temporal as well as spatial mapping. One of the functions of the temporal map would be to tune out irrelevant stimuli. Although we have considered elsewhere the possibility that the time at which an episode occurred might be indirectly coded in the mapping system (see O'Keefe & Nadel, 1978), this is a different type of temporal coding from the short-term temporal sequencing between stimuli invoked by Solomon.

We have looked carefully at the conditioning literature that Solomon cited and concluded that with the exception of blocking, the cognitive map theory handles the results adequately. Whether or not it can also handle blocking must await experiments designed to elucidate the role of place learning in this paradigm.

Reference Notes

1. Kohler, E. A., & Ayres, J. J. B. *The Kamin blocking effect with variable duration CSs*. Paper presented at the meeting of the Eastern Psychological Association, Washington, D. C., April 1978.
2. Nadel, L., & Willner, J. *The hippocampus and classical conditioning*. Manuscript in preparation, 1979.
3. O'Keefe, J. *A review of the hippocampal place cells*. Manuscript submitted for publication, 1979.

References

Ackil, J. E., Mellgren, R. L., Halgren, C., & Frommer, G. P. Effects of CS preexposures on avoidance learning in rats with hippocampal lesions. *Journal of Comparative and Physiological Psychology*, 1969, *69*, 739–747.

Berger, T. W., Alger, B. E., & Thompson, R. F. Neuronal substrates of classical conditioning in the hippocampus. *Science*, 1976, *192*, 483–485.

Berger, T. W., & Thompson, R. F. Identification of pyramidal cells as the central elements in hippocampal neuronal plasticity during learning. *Proceedings of the National Academy of Science*, 1978, *75*, 1572–1576. (a)

Berger, T. W., & Thompson, R. F. Neuronal plasticity in the limbic system during classical conditioning of the rabbit's nictitating membrane response: 1. The hippocampus. *Brain Research*, 1978, *145*, 323–336. (b)

Best, P. J., & Ranck, J. B., Jr. Reliability of the relationship between hippocampal unit activity and behavior in the rat. *Neuroscience Abstracts I*, 1975, *1*, 538.

Black, A. H., Nadel, L., & O'Keefe, J. Hippocampal function in avoidance learning and punishment. *Psychological Bulletin*, 1977, *84*, 1107–1129.

Dickinson, A., Hall, G., & Mackintosh, N. J. Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, 1976, *2*, 313–322.

Dweck, C. S., & Wagner, A. R. Situational cues and correlation between CS and US as determinants of the conditioned emotional response. *Psychonomic Science*, 1970, *18*, 145–147.

Elliott, K., & Whelan, J. (Eds.). *Functions of the septo-hippocampal system*. Amsterdam: Elsevier, 1978.

Fox, S. E., & Ranck, J. B., Jr. Localization and anatomical identification of theta and complex spike cells in dorsal hippocampal formation of rats. *Experimental Neurology*, 1975, *49*, 299–313.

Fox, S. E., & Ranck, J. B., Jr. Hippocampal complex spike and theta cell activity evoked by stimulation of limbic structures in unrestrained rats. *Neuroscience Abstracts III*, 1977, *3*, 198.

Goddard, G. V., McIntyre, D. C., & Leech, C. K. A permanent change in brain function resulting from daily electrical stimulation. *Experimental Neurology*, 1969, *25*, 295–330.

Gray, T. Blocking in the CER: Trace and delay procedures. *Canadian Journal of Psychology*, 1978, *32*, 40–42.

Gray, T., & Appignanesi, A. A. Compound conditioning: Elimination of the blocking effect. *Learning and Motivation*, 1973, *4*, 374–380.

Harper, R. M. Frequency changes in hippocampal electrical activity during movement and tonic immobility. *Physiology and Behavior*, 1971, *7*, 55–58.

Hill, A. J. First occurrence of hippocampal spatial firing in a new environment. *Experimental Neurology*, 1978, *62*, 282–297.

Kamin, L. J. Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior*. New York: Appleton-Century-Crofts, 1969.

Kramis, R., Vanderwolf, C. H., & Bland, B. H. Two types of hippocampal rhythmical slow activity in both the rabbit and the rat: Relations to behavior and effects of atropine, diethyl ether, urethane and pentobarbital. *Experimental Neurology*, 1975, *49*, 58–85.

Lubow, R. E. Latent inhibition. *Psychological Bulletin*, 1973, *79*, 398–407.

Lubow, R. E., Rifkin, B., & Alek, M. The context effect: The relationship between stimulus preexposure

and environmental preexposure determines subsequent learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 1976, *2*, 38–47.
Mackintosh, N. J. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 1975, *82*, 276–298.
McFarland, D. J., Kostas, J., & Drew, W. G. Dorsal hippocampal lesions: Effects of preconditioning CS exposure on flavor aversion. *Behavioral Biology*, 1978, *22*, 398–404.
Micco, D. J., & Schwartz, M. Effects of hippocampal lesions upon the development of Pavlovian internal inhibition in rats. *Journal of Comparative and Physiological Psychology*, 1971, *76*, 371–377.
Nadel, L., & O'Keefe, J. The hippocampus in pieces and patches: An essay on modes of explanation in physiological psychology. In R. Bellairs & E. G. Gray (Eds.), *Essays on the nervous system: A festschrift for Professor J. Z. Young*. Oxford: The Clarendon Press, 1974.
Nadel, L., O'Keefe, J., & Black, A. H. Slam on the brakes: A critique of Altman, Brunner and Bayer's response inhibition model of hippocampal function. *Behavioral Biology*, 1975, *14*, 151–162.
Odling-Smee, F. J. The role of background stimuli during Pavlovian conditioning. *Quarterly Journal of Experimental Psychology*, 1975, *27*, 201–209.
Odling-Smee, F. J. The over-shadowing of background stimuli: Some effects of varying amounts of training and UCS intensity. *Quarterly Journal of Experimental Psychology*, 1978, *30*, 737–746.
O'Keefe, J. Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, 1976, *51*, 78–109.
O'Keefe, J., & Conway, D. H. Hippocampal place units in the freely moving rat: Why they fire where they fire. *Experimental Brain Research*, 1978, *31*, 573–590.
O'Keefe, J., & Dostrovsky, J. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely moving rat. *Brain Research*, 1971, *34*, 171–175.
O'Keefe, J. & Nadel, L. *The hippocampus as a cognitive map*. Oxford: Clarendon Press, 1978.
Olton, D., Branch, M., & Best, P. Spatial correlates of hippocampal unit activity. *Experimental Neurology*, 1978, *58*, 387–409.
Olton, D. S., & Isaacson, R. L. Hippocampal lesions and active avoidance. *Physiology and Behavior*, 1968, *3*, 719–724.
Pavlov, I. P. *Conditioned reflexes*. London: Oxford University Press, 1927.
Racine, R. J. Modification of seizure activity by electrical stimulation: 1. After-discharge. *Electroencephalography and Clinical Neurophysiology*, 1972, *32*, 269–279.
Rescorla, R. A. Pavlovian conditioned inhibition. *Psychological Bulletin*, 1969, *72*, 77–94.
Rescorla, R. A., & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II*. New York: Appleton-Century-Crofts, 1972.
Rickert, E. J., Bennett, T. L., Lane, P., & French, J. Hippocampectomy and attenuation of blocking. *Behavioral Biology*, 1978, *22*, 147–160.
Salafia, W. R., Chiaia, N. L., & Ramirez, J. J. Retardation of rabbit nictitating membrane response conditioning by sub-seizure electrical stimulation of the hippocampus. *Physiology and Behavior*, 1979, *22*, 451–455.
Salafia, W. R., Romano, A. G., Tynan, T. T., & Host, K. C. Disruption of rabbit (*Oryctolagus cuniculus*) nictitating membrane conditioning by post-trial electrical stimulation of the hippocampus. *Physiology and Behavior*, 1977, *18*, 207–212.
Sheafor, P. J. "Pseudoconditioned" jaw movements of the rabbit reflect associations conditioned to contextual background cues. *Journal of Experimental Psychology: Animal Behavior Processes*, 1975, *104*, 245–260.
Sheffield, F. Relation between classical conditioning and instrumental learning. In W. F. Prokasy (Ed.), *Classical conditioning: A symposium*. New York: Appleton-Century-Crofts, 1965.
Solomon, P. R. Role of the hippocampus in blocking and conditioned inhibition of the rabbit's nictitating membrane response. *Journal of Comparative and Physiological Psychology*, 1977, *91*, 407–417.
Solomon, P. R. Temporal versus spatial information processing theories of hippocampal function. *Psychological Bulletin*, 1979, *86*, 1272–1279.
Solomon, P. R., & Moore, J. W. Latent inhibition and stimulus generalization of the classically conditioned nictitating membrane response in rabbits (*Oryctolagus cuniculus*) following dorsal hippocampal ablation. *Journal of Comparative and Physiological Psychology*, 1975, *89*, 1192–1203.
Thompson, R. F. The search for the engram. *American Psychologist*, 1976, *31*, 209–227.
Vinogradova, O. Registration of information and the limbic system. In G. Horn & R. A. Hinde (Eds.), *Short term changes in neural activity and behavior*. London: Cambridge University Press, 1970.
Vinogradova, O. S. Functional organization of the limbic system in the process of registration of information: Facts and hypotheses. In R. L. Isaacson & K. H. Pribram (Eds.), *The hippocampus*. New York: Plenum Press, 1975.
Wagner, A. R., & Rescorla, R. A. Inhibition in Pavlovian conditioning: Application of a theory. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and learning*. London: Academic Press, 1972.
Winson, J. Inter-species differences in the occurrence of theta. *Behavioral Biology*, 1972, *7*, 479–487.
Zener, K. The significance of behavior accompanying conditioned salivary secretion for theories of the conditioned response. *American Journal of Psychology*, 1937, *50*, 384–403.

Received May 20, 1979 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1290–1297

Type I Error Rate of the Chi-Square Test of Independence in $R \times C$ Tables That Have Small Expected Frequencies

Drake R. Bradley, T. D. Bradley, Steven G. McGrath, and Steven D. Cutcomb
Bates College

Sampling experiments are reported which show that the uncorrected chi-square test of independence is exceptionally robust with respect to small expected frequencies in $R \times C$ contingency tables. In 2×2, 2×3, 3×3, 3×4, and 4×4 tables, the actual Type I error rates did not exceed .06 ($\alpha = .05$) for those applications most likely to arise in practice. In general, error rates that exceeded .06 occurred only when both marginal probability distributions were extremely skewed and sample size was small. Nevertheless, the quality of the approximation of chi-square probabilities for exact multinomial probabilities was sometimes poor, although excessive errors in approximation by Cochran's criteria usually resulted from actual error rates being smaller, not larger, than the nominal level. A distinction is made between accuracy of approximation and control of the Type I error rate as considerations dictating the advisability of using an approximate test.

In 1949 Lewis and Burke published an article entitled "The Use and Misuse of the Chi-Square Test." These authors argued that social scientists frequently misused the chi-square test and that the most common error in applying the test to contingency tables resulted from "the use of extremely small theoretical frequencies" (Lewis & Burke, 1949, p. 460). Rebuttals by Peters (1950), Pastore (1950), and Edwards (1950) did little to resolve the developing controversy, and recent publications (Tate & Hyer, 1973; Tate & Hyer, Note 1) demonstrate that the issue of "how small is small?" is still unresolved today. The issue arises, of course, because the chi-square distribution provides only approximate estimates of exact multinomial probabilities. The quality of this approximation depends in a complex way on sample size, the true distribution of marginal probabilities in the population, the number of cells in the contingency table, and the level of significance employed. Consequently, it is extremely difficult to formulate simple rules of thumb that indicate when the approximation is adequate and when it is not.

Cochran (1952) suggested that the approximation is satisfactory if the actual Type I error probabilities remain between .04 and .06 for tests conducted at $\alpha = .05$ or between .007 and .015 for tests conducted at $\alpha = .01$. Given any such specific criteria, the traditional approach has been to seek a minimum expectation that delimits the lower bound beyond which the approximation is no longer satisfactory. Although there is no commonly accepted rule of thumb for deciding when expected frequencies are too small, many investigators accept Hays's recommendation (1963, p. 584, 597) that for applications involving one degree of freedom, the minimum should be set at 10, whereas for other applications a minimum of 5 is sufficient. Under some circumstances a minimum expectation of 2, 1, or even fractional expectations may be permissible according to some authors (Cochran, 1952, 1954; Good, 1961; Slakter, 1965, 1966;

The authors wish to acknowledge the excellent service provided by the Dartmouth Time-Sharing system, which was used to conduct the computer simulations reported in this article.

Requests for reprints should be sent to Drake R. Bradley, Department of Psychology, Bates College, Lewiston, Maine 04240.

Vessereau, 1958; Wise, 1963; Yarnold, 1970). However, it has been recently suggested that a minimum expectation of 20 is necessary to ensure that the approximation will be accurate under all circumstances (Tate & Hyer, 1973). Unfortunately, the absence of a strong consensus on this matter makes it extremely difficult for the researcher to know if he or she is employing the chi-square statistic properly.

A related issue concerns the use of a continuity or similar correction to improve the approximation. There has been a continuing controversy over whether such corrections are necessary or desirable in all instances, and assuming that they are, over which method of correction provides the best approximation (Boschloo, 1970; Camilli & Hopkins, 1978; Conover, 1974; Garside, 1971, 1972; Garside & Mack, 1970, 1976; Grizzel, 1967; Larntz, 1978; Mantel, 1974; Mantel & Greenhouse, 1968; Miettinen, 1974; Nass, 1959; Plackett, 1964; Starmer, Grizzel, & Sen, 1974; Yates, 1934). Part of the controversy results from differences in opinion as to what constitutes a good approximation. Some authors insist that whatever correction is used, the actual Type I error probability should be maintained at a level less than or equal to α (Mantel & Greenhouse, 1968). Others argue that the correction should simply minimize the absolute difference between the true and nominal values. The latter technique, although it permits some inflation in the Type I error rate, may often be more accurate than the former. There has also been disagreement as to whether so-called exact tests provide the appropriate standard for evaluating the ability of approximate tests to control the Type I error rate at the nominal level (Starmer et al., 1974). This issue arises because the exact test may not provide a test with a predetermined significance level as high as the desired nominal value (Garside & Mack, 1976; Starmer et al., 1974) unless it is supplemented by randomization techniques (Tocher, 1950). If so, continuity corrections based on the exact test as the standard may be unnecessarily conservative.

Indeed, a number of Monte Carlo studies have shown that the chi-square test is relatively robust with regard to violations of the minimum expected frequency requirement, even when the test is uncorrected. Slakter (1966) found that the chi-square test was robust in goodness of fit to uniform applications even with fractional expected frequencies. Lewontin and Felsenstein (1965) investigated the $2 \times K$ case in a uniform population and where both row and column marginal frequencies remained fixed and found that the chi-square test was robust over the range of values tested. Roscoe and Byars (1971) extended Slakter's findings to a wider range of N (sample size) and K (number of categories) as well as to nonuniform distributions. In addition, they extended Lewontin and Felsenstein's results to contingency tables that had rows and columns in all combinations of 2 to 5, although they elected to fix only the row marginal frequencies. They found that the chi-square test was exceptionally robust in goodness of fit to uniform applications and in tests of independence with more than one degree of freedom. Finally, Bradley and Cutcomb (1977) and Camilli and Hopkins (1978) have shown that for selected sets of 2×2 tables in which neither row nor column marginals were fixed, the chi-square test of independence is highly robust to violations of minimum expected cell frequency. The results of these various Monte Carlo studies suggest that the minimum expected frequency requirement can be relaxed considerably.

In this article we report the results of a comprehensive set of sampling experiments that evaluate the Type I error rate of the chi-square test of independence in $R \times C$ contingency tables. The simulations reported here differ from previous numerical or Monte Carlo studies on $R \times C$ contingency tables (Camilli & Hopkins, 1978; Garside & Mack, 1976; Kurtz, 1968; Larntz, 1978; Lewontin & Felsenstein, 1965; Roscoe & Byars, 1971; Starmer et al., 1974) in that (a) the sample sizes and marginal probability distributions that were selected range over all values likely to arise in practice, and/or (b) neither the row nor column marginal frequencies were fixed. The second feature is intended to model the most common application of chi-square to $R \times C$ tables; that is, N elements are randomly sampled from a population and classified on each of two polychotomous variables, thereby producing a table in which the row and

Table 1
Marginal Probability Distributions in Which One or More Type I Error Rates Were Outside the Interval .03–.06 for the Chi-Square Test of Independence in R × C *Tables*

Row	Marginal probabilities: Row 1	Row 2	Row 3	Row 4	Column 1	Column 2	Column 3	Column 4	Sample size: 20	30	40	60	80	100	200
								$p > .06$							
1.	.9	.1	—	—	.9	.1	—	—	.0721	—	.0568	.0492	.0437	.0425	.0412
2.	.9	.1	—	—	.8	.1	.1	—	.0715	—	.0667	.0581	.0511	.0486	.0410
3.	.7	.2	.1	—	.8	.1	.1	—	.0718	—	.0604	.0580	.0534	.0514	.0504
4.	.8	.1	.1	—	.8	.1	.1	—	.0891	—	.0820	.0667	.0582	.0543	.0515
5.	.7	.2	.1	—	.7	.1	.1	.1	.0578	.0657	.0635	.0578	.0552	.0557	.0481
6.	.8	.1	.1	—	.4	.4	.1	.1	.0472	.0584	.0609	.0516	.0516	.0518	.0460
7.	.8	.1	.1	—	.5	.3	.1	.1	.0510	.0648	.0607	.0609	.0554	.0540	.0494
8.	.8	.1	.1	—	.6	.2	.1	.1	.0588	.0643	.0681	.0601	.0533	.0495	.0482
9.	.8	.1	.1	—	.7	.1	.1	.1	.0666	.0772	.0764	.0642	.0596	.0589	.0515
10.	.4	.4	.1	.1	.7	.1	.1	.1	.0471	.0604	.0605	.0539	.0591	.0550	.0478
11.	.5	.3	.1	.1	.7	.1	.1	.1	.0474	.0603	.0606	.0509	.0558	.0500	.0529
12.	.6	.2	.1	.1	.6	.2	.1	.1	.0446	.0544	.0615	.0542	.0489	.0500	.0531
13.	.6	.2	.1	.1	.7	.1	.1	.1	.0537	.0626	.0602	.0572	.0550	.0546	.0499
14.	.7	.1	.1	.1	.7	.1	.1	.1	.0581	.0746	.0755	.0668	.0577	.0573	.0503

Table 1 (*continued*)

	Marginal probabilities								Sample size						
	Row				Column										
Row	1	2	3	4	1	2	3	4	20	30	40	60	80	100	200
							$p < .03$								
15.	.5	.5	—	—	.9	.1	—	—	.0211	—	.0359	.0505	.0477	.0520	.0473
16.	.6	.4	—	—	.9	.1	—	—	.0287	—	.0390	.0418	.0477	.0528	.0504
17.	.5	.5	—	—	.8	.1	.1	—	.0187	—	.0327	.0419	.0480	.0486	.0509
18.	.6	.4	—	—	.8	.1	.1	—	.0227	—	.0383	.0397	.0427	.0470	.0502
19.	.9	.1	—	—	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	—	.0240	—	.0371	.0389	.0475	.0447	.0479
20.	.9	.1	—	—	.4	.3	.3	—	.0274	—	.0357	.0410	.0444	.0438	.0474
21.	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	—	.7	.2	.1	—	.0277	—	.0424	.0458	.0450	.0444	.0468
22.	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	—	.8	.1	.1	—	.0229	—	.0324	.0402	.0399	.0433	.0491
23.	.4	.3	.3	—	.8	.1	.1	—	.0209	—	.0350	.0372	.0474	.0453	.0513
24.	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	—	.4	.4	.1	.1	.0266	.0353	.0393	.0409	.0433	.0477	.0507
25.	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	—	.5	.3	.1	.1	.0251	.0345	.0391	.0398	.0449	.0504	.0473
26.	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	—	.6	.2	.1	.1	.0258	.0308	.0358	.0445	.0448	.0465	.0504
27.	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	—	.7	.1	.1	.1	.0201	.0249	.0339	.0393	.0422	.0412	.0507
28.	.4	.3	.3	—	.4	.4	.1	.1	.0278	.0362	.0405	.0404	.0442	.0462	.0488
29.	.4	.3	.3	—	.5	.3	.1	.1	.0297	.0342	.0429	.0468	.0465	.0499	.0490
30.	.4	.3	.3	—	.6	.2	.1	.1	.0266	.0353	.0397	.0443	.0403	.0466	.0459
31.	.4	.3	.3	—	.7	.1	.1	.1	.0211	.0297	.0341	.0404	.0445	.0445	.0499
32.	.4	.4	.2	—	.7	.1	.1	.1	.0272	.0352	.0389	.0408	.0469	.0451	.0512
33.	.8	.1	.1	—	.25	.25	.25	.25	.0276	.0317	.0356	.0396	.0400	.0426	.0445
34.	.25	.25	.25	.25	.4	.4	.1	.1	.0291	.0366	.0397	.0409	.0442	.0492	.0512
35.	.25	.25	.25	.25	.5	.3	.1	.1	.0289	.0313	.0401	.0459	.0440	.0474	.0470
36.	.25	.25	.25	.25	.7	.1	.1	.1	.0266	.0306	.0350	.0394	.0455	.0445	.0468
37.	.3	.3	.2	.2	.4	.4	.1	.1	.0287	.0352	.0388	.0424	.0431	.0429	.0512
38.	.3	.3	.2	.2	.6	.2	.1	.1	.0266	.0352	.0386	.0418	.0439	.0519	.0439

Note. $\alpha = .05$, $N \geq 20$.

column marginals may vary from one sample to the next.

Method

Actual Type I error rates for the chi-square test in 2×2, 2×3, 3×3, 3×4, and 4×4 tables were determined by sampling experiments conducted on a Honeywell 66/40 duplex computer. For any given simulation, the joint probability distribution was obtained by multiplying row and column marginal probabilities: $P(RC_{ij}) = P(R_i)P(C_j)$. A pseudorandom number generator (linear congruence method) sampled N numbers between 0 and 1 from a uniform distribution, and each such number was used to place a count in one of the cells of the $R \times C$ table via reference to the cumulative joint probability distribution (see Lehman, 1977, p. 166). The chi-square test of independence (Hays, 1963, p. 591) was conducted on the resulting contingency table, using the empirically sampled marginal frequencies to compute the expected frequency of each cell.[1] This procedure was repeated for a total of 10,000 samples per simulation, and the proportion of significant chi-square tests that were obtained provided the estimate of the Type I error rate.

Each $R \times C$ table was investigated for a large number of marginal probability distributions ranging from uniform to highly skewed on both rows and columns and including all combinations (in .10 increments) in between. For example, 2×2 tables were investigated for row and column marginal probabilities ranging from .5, .5 and .5, .5 (uniform on both) to .9, .1 and .9, .1 (highly skewed on both). All combinations in between these extremes, in .10 increments (.5, .5 and .6, .4; .5, .5 and .7, .3; etc.), were investigated. Similarly, in 3×4 tables the marginal probabilities ranged from $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$, and .25, .25, .25, .25 (uniform on both) to .8, .1, .1 and .7, .1, .1, .1 (highly skewed on both). Every $R \times C$ marginal probability distribution was investigated at $\alpha = .05$ and $N = 20$, 40, 60, 80, 100, and 200. In addition, all 2×3 and 3×3 tables were investigated for $N = 10$, all 3×4 and 4×4 tables were investigated for $N = 30$, and all 2×2 tables were investigated for $N = 4$, 10, 400, and 1000. Finally, all 2×2 tables were also investigated at the $\alpha = .01$ and .10 levels of significance.

Fifteen combinations of marginal probabilities were investigated for 2×2 tables, 45 for 2×3 tables, 45 for 3×3 tables, 90 for 3×4 tables, and 55 for 4×4 tables. Given a grand total of 250 unique $R \times C$ marginal probability distributions and the various sample sizes and α levels that were investigated, 2095 individual simulations were conducted to evaluate the actual Type I error rates of the chi-square statistic in $R \times C$ contingency tables.

Results

Due to space limitations, complete tables of the Type I error rate of chi-square in 2×2, 2×3, 3×3, 3×4, and 4×4 tables cannot be presented here.[2] Table 1 presents Type I error rates for those marginal probability distributions in which one or more error rates fall outside the interval .03–.06 ($N \geq 20$ and $\alpha = .05$).[3] Hence, Table 1 summarizes those conditions in which the errors of approximation might be considered excessive. Note that we have not employed Cochran's (1952) suggested interval of .04–.06 in constructing Table 1. Although the use of an asymmetrical interval (.03–.06) was motivated partly by the requirement to present the data in highly condensed form (errors of approximation in the .03–.04 region were fairly common for $N = 20$), the asymmetry also parallels that implicitly endorsed by most investigators, namely, that errors of approximation of any given magnitude are generally more tolerable when they err in the direction of conservatism.[4]

---

[1] On occasion, a sampled table might contain one or more expected frequencies equal to zero; if so, the corresponding cells were bypassed in the computation of the total value of chi-square to prevent division by zero.

[2] Some of the data discussed in this report are reproduced in Bradley and Cutcomb (1977) and Bradley, McGrath, and Bradley (1978), and were presented at the 1977 Eastern Regional Meeting in Chapel Hill, North Carolina, and the 1978 Annual Meeting in San Diego, California, of the American Statistical Association.

Unabbreviated tables of Type I error rates are available from the first author.

[3] As a check on the Monte Carlo procedure used, the empirical estimates can be compared to the exact probabilities of a Type I error computed for 2×2 tables based on small N. Consider, for example, an N of 4 and a marginal probability distribution of .5, .5 and .5, .5: manual expansion of the multinominal results in an exact probability of .1094 of sampling a table which will produce a chi-square statistic significant at the .05 level or better. The corresponding sampling experiment produced an empirical estimate of .1089, which agrees with the exact value within the limits of error ($\sigma_p = \sqrt{p(1-p)/K}$, where $K = 10{,}000$ trials). Similar checks for the remaining 2×2 marginal probability distributions that were investigated showed the empirical estimates to be in good agreement with the exact values.

[4] Of course, the investigator must also be concerned with maintaining adequate power in conducting tests of independence in situations in which errors of approximation may result in differences between actual and nominal power. Bradley and Seely (1977) have developed power tables for the 2×2 case that because they are based on empirical simulations, automatically incorporate the effect of errors of approximation on the power of the chi-square test.

The increased probability of making a Type I error is usually of greatest concern, and for this reason we have used the same upper limit (.06) for our interval as that used by Cochran (1952).

Several trends may be noted with respect to the data presented in Table 1. First, excessive inflation in the Type I error rate ($p > .06$) tends to be a problem when both marginal probability distributions of the $R \times C$ table are highly skewed (e.g., rows 1, 2, 4, 9, and 14). Conversely, excessive deflation in the Type I error rate ($p < .03$) tends to be a problem when one marginal distribution of the $R \times C$ table is highly skewed and the other is relatively uniform (e.g., rows 15, 17, 22, 27, and 36). Finally, marginal probability distributions that are uniform (or nearly uniform) on both rows and columns do not result in excessive errors of approximation, even when minimum expected frequency requirements (of 5 or 10) are violated in all $R \times C$ cells of the table, as often occurs for $N = 20$. [Theoretical expected frequencies may be obtained by computing $P(R_i)P(C_j)N$ for each cell of the table.]

The error rates represented in Table 1 pertain to tests conducted at $\alpha = .05$. As noted earlier, simulations were also conducted for $\alpha = .01$ and .10 for 2×2 tables. Although the trends discussed previously were clearly evident in the error rate data for $\alpha = .01$ ($N \geq 20$), they were less apparent in the data for $\alpha = .10$.[5]

Discussion

These results might or might not imply a "liberalized" policy with regard to using the chi-square test on tables that have small expected frequencies. Since errors of approximation can be fairly substantial, with actual error rates ranging from .0187 to .0891 for $N \geq 20$ and $\alpha = .05$ (rows 17 and 4 of Table 1), some investigators might elect to maintain a conservative policy with respect to using the chi-square test (see Tate & Hyer, 1973). However, traditional rules of thumb based on minimum expected frequency, without regard to the marginal distributions, do not provide selective protection against errors of approximation where such protection is needed most (Table 1). That is, these rules of thumb will often prohibit the use of the chi-square test in situations in which it provides a satisfactory approximation. Furthermore, it can be argued that accuracy of approximation per se is not the central issue in determining the advisability of using the chi-square test. Rather, the key issue for many investigators is the ability of this test to control the Type I error rate at or below some acceptable upper limit, α', relative to the nominal level, α.

The sampling experiments reported here show that for the large majority of applications likely to arise in practice, the actual Type I error rates will not exceed $\alpha' = .06$ for tests conducted at the nominal level of $\alpha = .05$.[6] This is true without any correction for continuity and regardless of the size and number of small expected frequencies in the $R \times C$ table. For the few exceptions that are noted in Table 1 ($p > .06$), in which row and column probabilities are both highly skewed, the investigator can simply select a more conservative alpha level for conducting the test. Since the maximum inflation in error rate for $N \geq 20$ is less than double the nominal value, using an adjusted level of $\alpha/2$ should suffice in most cases for the provision of adequate protection at the original level desired.

The preceding arguments do not, of course, imply that when the investigator consults a table of chi-square to determine the specific level of significance of an outcome, the tabled value will provide an accurate approximation to the exact multinomial probability. Tate and Hyer (1973) are correct in noting that these values can differ substantially. Never-

[5] To investigate in more detail the effect of small sample size on errors of approximation, additional sampling experiments were conducted at $\alpha = .05$ for a 2×2 table with marginal probabilities of .8, .2 and .9, .1. Actual Type I error rates were obtained for all sample sizes between and including $N = 4$ and $N = 40$. The Type I error rates ranged from .0203 ($N = 7$) to .0589 ($N = 20$).

[6] For $\alpha = .01$ (2×2 tables) a parallel conclusion is reached, namely, that for the large majority of applications likely to arise in practice, the actual Type I error rates will not exceed $\alpha' = .015$ (Cochran's upper limit) for tests conducted at the $\alpha = .01$ level. As in Table 1 (row 1), the only exception to this generalization was the .9, .1 and .9, .1 marginal probability distribution.

theless, as long as the null hypothesis can be tested at some maximum specifiable level of risk, inaccuracy of approximation on either side of the rejection point need not concern the investigator, if his or her only interest is in making a decision about the presence of association in the $R \times C$ table and not in estimating exact cumulative multinomial probabilities. In this context, it seems unreasonable to prohibit the use of a convenient statistical test on the basis of inaccuracies in approximation over regions of the distributions that are largely irrelevant to the decision-making behavior of the investigator.

In conclusion, if the investigator's main concern is in controlling the Type I error rate or below some acceptable upper limit ($\alpha' = .06$), then the computer simulations reported in this article show that the chi-square test of independence may be safely used in nearly all situations likely to arise in practice. If this conclusion is correct, then the frequent abuse by social scientists of the minimum expected frequency requirement that was first reported by Lewis and Burke in 1949 will turn out not to be a "misuse" of the chi-square test at all. Instead, it will simply illustrate another instance in which "the results of theory sometimes remain substantially true even when some assumptions fail to hold" (Cochran, 1952, p. 328).

Reference Note

1. Tate, M. W., & Hyer, L. A. *Significance values for an exact multinomial test and accuracy of the chi-square approximation* (Report No. ED 040 886). Bethlehem, Pa.: Lehigh University, August 1969.

References

Boschloo, R. D. Raised conditional level of significance for the 2×2 table when testing for the equality of two probabilities. *Statistica Neerlandica*, 1970, *21*, 1–35.

Bradley, D. R., & Cutcomb, S. Monte Carlo simulations and the chi-square test of independence. *Behavior Research Methods and Instrumentation*, 1977, *9*, 193–201.

Bradley, D. R., McGrath, S. G., & Bradley, Lt. Col. T. D. The Type I error rate of the chi-square test of independence in $R \times C$ contingency tables. *Proceedings of the Statistical Computing Section of the American Statistical Association*, 1978, 212–217.

Bradley, D. R., & Seely, D. L. Empirical determination of the power of the chi-square test of independence in 2×2 tables. *Proceedings of the Statistical Computing Section of the American Statistical Association*, 1977, 138–144.

Camilli, G., & Hopkins, K. D. Applicability of chi-square to 2×2 contingency tables with small expected cell frequencies. *Psychological Bulletin*, 1978, *85*, 163–167.

Cochran, W. G. The X^2 test of goodness of fit. *Annals of Mathematical Statistics*, 1952, *23*, 315–345.

Cochran, W. G. Some methods for strengthening the common χ^2 tests. *Biometrics*, 1954, *10*, 417–451.

Conover, W. J. Some reasons for not using the Yates continuity correction on 2×2 contingency tables. *Journal of the American Statistical Association*, 1974, *69*, 374–376.

Edwards, A. L. On the use and misuse of the chi-square test—The case of the 2×2 contingency table. *Psychological Bulletin*, 1950, *47*, 341–346.

Garside, G. R. An accurate correction for the χ^2 test in the homogeneity case of 2×2 contingency tables. *New Journal of Statistics and Operational Research*, 1971, *7*, 1–26.

Garside, G. R. Further tables of an accurate correction for the χ^2-test in the homogeneity case of the 2×2 contingency table. *New Journal of Statistics and Operational Research*, 1972, *8*, 6–25.

Garside, G. R., & Mack, C. A quantitative analysis of all sources of correction in the homogeneity case of the 2×2 contingency table. *New Journal of Statistics and Operational Research*, 1970, *6*, 16–25.

Garside, G. R., & Mack, C. Actual Type I error probabilities for various tests in the homogeneity case of the 2×2 contingency table. *The American Statistician*, 1976, *30*, 18–21.

Good, I. J. The multivariate saddlepoint method and chi-squared for the multinomial distribution. *Annals of Mathematical Statistics*, 1961, *32*, 535–548.

Grizzel, J. Continuity correction in the χ^2-test for 2×2 tables. *The American Statistician*, 1967, *21*, 28–32.

Hays, W. L. *Statistics*. New York: Holt, Rinehart & Winston, 1963.

Kurtz, T. E. A role of time-sharing computing in statistical research. *The American Statistician*, 1968, *22*, 19–21.

Larntz, K. Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, 1978, *73*, 253–263.

Lehman, R. S. *Computer simulation and modeling*. New York: Wiley, 1977.

Lewis, D., & Burke, C. J. The use and misuse of the chi-square test. *Psychological Bulletin*, 1949, *46*, 433–489.

Lewontin, R. C., & Felsenstein, J. The robustness of homogeneity tests in $2 \times N$ tables. *Biometrics*, 1965, *21*, 19–33.

Mantel, N. Comment and a suggestion. *Journal of the American Statistical Association*, 1974, *69*, 378–380.

Mantel, N., & Greenhouse, S. W. What is the continuity correction? *The American Statistician*, 1968, *22*, 27–30.

Miettinen, O. S., Comment. *Journal of the American Statistical Association*, 1974, *69*, 380–382.
Nass, C. A. G. The χ^2 test for small expectations in contingency tables with special reference to accidents and absenteeism. *Biometrika*, 1959, *46*, 365–385.
Pastore, N. Some comments on the use and misuse of the chi-square test. *Psychological Bulletin*, 1950, *47*, 338–340.
Peters, C. C. The misuse of chi-square–A reply to Lewis and Burke. *Psychological Bulletin*, 1950, *47*, 331–337.
Plackett, R. L. The continuity correction in 2×2 tables. *Biometrika*, 1964, *51*, 327–337.
Roscoe, J. T., & Byars, J. A. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association*, 1971, *66*, 755–759.
Slakter, M. J. A comparison of the Pearson chi-square and Kolmogorov goodness-of-fit test with respect to validity. *Journal of the American Statistical Association*, 1965, *60*, 854–858
Slakter, M. J. Comparative validity of the chi-square and two modified chi-square goodness-of-fit tests for small but equal expected frequencies. *Biometrika*, 1966, *53*, 619–622.
Starmer, C., Grizzel, J. E., & Sen, P. K. Comment. *Journal of the American Statistical Association*, 1974, *69*, 376–378.
Tate, M. W., & Hyer, L. A. Inaccuracy of the X^2 test of goodness-of-fit when expected frequencies are small. *Journal of the American Statistical Association*, 1973, *68*, 836–841.
Tocher, K. D. Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika*, 1950, *37*, 130–144.
Vessereau, A. [Sur les conditions d'application du criterium χ^2 de Pearson.] *Bulletin de l'Insitut International de Statistique*, 1958, *36*, 87–101. (*Mathematical Reviews*, 1961, *22*, No. 10069. [Abstract]
Wise, M. E. Multinomial probabilities and the X^2 and χ^2 distributions. *Biometrika*, 1963, *50*, 145–154.
Yarnold, J. K. The minimum expectation in X^2 goodness-of-fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, 1970, *65*, 864–886.
Yates, F. Contingency tables involving small numbers and the χ^2-test. *Journal of the Royal Statistical Society Supplement*, 1934, *1*, 217–235.

Received July 3, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1298–1308

Home-Based Reinforcement Programs Designed to Modify Classroom Behavior: A Review and Methodological Evaluation

Beverly M. Atkeson and Rex Forehand
University of Georgia

The present article reviews and evaluates the current research on the use of home-based reinforcement programs to modify both disruptive and academic behaviors in the classroom. This treatment program has been used with a variety of students, settings, and behaviors, and the results have been positive. However, the conclusions concerning the effectiveness of home-based reinforcement programs are limited because of the methodologies used in the studies that are reviewed.

Over the past 10 to 15 years, the role of the behavioral psychologist who works with children has gradually changed from that of direct change agent to that of consultant. More and more frequently, parents are trained as behavior therapists for their child's problems. Recent reviews indicate that this is an effective and efficient treatment approach (Berkowitz and Graziano, 1972; Graziano, 1977; Johnson & Katz, 1973; O'Dell, 1974). In the field of education, a similar shift has occurred. There is now substantial experimental evidence which indicates that teachers themselves can function as behavior modifiers in the classroom and can deal with a wide variety of problems (O'Leary & O'Leary, 1976, 1977b).

More recently, an increasing number of studies have tried to join the teacher and parent in a cooperative effort to reduce problems in the school setting. The present article reviews the current research on the use of home-based reinforcement programs designed to modify classroom behaviors. Basically, with this procedure the teacher is responsible for specifying the classroom rules, for determining rule violations, and for communicating these to the parent. At home the parent is responsible for consistently dispensing rewards and sanctions to the child, based on the teacher's report.

For purposes of presentation, the articles reviewed have been organized into two areas, based on the classroom behaviors that were targeted for treatment. The first section reviews the research on home-based reinforcement programs designed to modify disruptive behaviors in the classroom, and the second section reviews those designed to change academic behaviors in the classroom. Following the presentation of the current research, the third section of the article examines the methodology used in the studies to evaluate the reported results and to determine future areas of research.

Disruptive Behaviors in the Classroom

Although specific behaviors labeled as disruptive may vary from teacher to teacher, in general they are behaviors that the teacher views as disruptive to the learning process in his or her classroom; these may include making noise, talking without permission, physically disturbing other children, and getting out of one's seat without permission.

This article is an expanded version of part of an article by Atkeson and Forehand entitled "Parents As Behavior Change Agents With School-Related Problems," which appeared in *Education and Urban Society* (*10*, 521–538).

Preparation of this manuscript was supported in part by National Institute of Mental Health Grant MH28859-01.

Requests for reprints should be addressed to Rex Forehand, Department of Psychology, University of Georgia, Athens, Georgia 30602.

Unfortunately, these behaviors often demand a great deal of the teacher's time. In fact, when teachers were asked to identify classroom behaviors to change, 80% chose to reduce the occurrence of a "bad" behavior (O'Leary & O'Leary, 1977a).

Teachers have successfully used a number of behavioral techniques to decrease disruptive behaviors in the classroom. These include (a) increasing incompatible behaviors (e.g., on-task behavior, sitting still, correct academic performance) through the use of contingent tangible rewards and/or social praise, (b) punishing disruptive behaviors (e.g., with loss of privileges and/or brief isolation of the child), and (c) dispensing tokens (e.g., check marks or chips) for following classroom rules or for meeting a specified level of academic performance, redeemable for backup reinforcers in the classroom (O'Leary & O'Leary, 1977b). Although highly effective, the above procedures do have some disadvantages. To be successful, each procedure may require a great deal of time and effort on the part of the teacher and behavioral consultant (Schumaker, Hovell, & Sherman, 1977). Often the teacher must alter her/his own teaching style to reduce the misbehavior of only a few students (Schumaker et al., 1977). In addition, the tangible rewards or backup reinforcers available are often limited in the classroom (Ayllon, Garber, & Pisor, 1975; Schumaker et al., 1977). In contrast, parents often have access to a wide variety of privileges (e.g., allowances, TV time, movies, skating) not available to most teachers (Ayllon et al., 1975; Bailey, Wolf, & Phillips, 1970; Karraker, 1972).

In an attempt to lessen the above disadvantages, psychologists have used some variation of a home-based reinforcement program to reduce the disruptive behavior of one or more children in a classroom. In these studies the content or degree of specificity of the teacher's report to the parents has varied from global to detailed. Some teachers were asked to send a note stating only whether the child was good (Ayllon et al., 1975; Heaton, Safer, Allen, Spinnato, & Prumo, 1976). More frequently, the note home listed the classroom rules and whether the child followed each of the rules (Hawkins, Sluyter, & Smith, 1972, Experiment 4; Lahey et al., 1977; Schumaker et al., 1977). Although both the global and specific reports to the home produced the desired change, there have been no studies that directly compared the two approaches. Each, of course, has its own advantages and disadvantages. The more global report may be somewhat quicker for the teacher to complete but may also be less objective. Stating each classroom rule on the note communicates more information to both the child and the parents. Three studies (Bailey et al., 1970; Clark, 1972; Kirigin, Bailey, Phillips, Fixen, & Wolf, Note 1) provided a compromise between global and detailed reports. At the beginning of the program, the teacher listed the classroom rules on the board so that they were clearly stated for the child. The daily report home, however, indicated only whether the student obeyed class rules.

In most studies, the teacher–parent communication occurred daily. Three studies, however, sent notes home only when the child was good (Ayllon et al., 1975; Hawkins et al., 1972, Experiment 4; Heaton et al., 1976), the rationale being that a note communicating misbehavior might never reach home (Ayllon et al., 1975). Only one study limited the parent–teacher communication to a weekly contact (Coleman, 1973). Although this was successful at maintaining the student's good behavior, the teacher had previously used other behavioral techniques to establish the desired behaviors in the classroom before instituting the weekly communication home. One would expect that the more frequent the communication between school and home, the more rapid the behavioral change.

Several studies did report that some students lost their notes before reaching home. The approach by Ayllon et al. (1975), Hawkins et al. (1972, Experiment 4), and Heaton et al. (1976) of sending notes home only when the report is good is one way to circumvent this problem, since such a procedure should increase the child's motivation for carrying the note home. However, this technique does not guarantee that a note will reach home. If the rewards dispensed at home are not potent enough, the note's importance to the child is diminished, and the note may be lost (Schumaker et al., 1977). Of course, the parents can be required to call the teacher to obtain

the necessary information if the note fails to reach home (Kroth, Whelan, & Stables, 1970). In fact, two studies used the telephone exclusively for the teacher–parent communications (Coleman, 1973; Strober & Bellack, 1975). Although this approach completely circumvents the problem of missing notes, it does require more time on the part of the teacher.

The rewards received by students at home for good behavior at school varied from verbal praise (Lahey et al., 1977; Schumaker et al., 1977, Experiment 2) to tangible rewards such as allowances (Coleman, 1973; Heaton et al., 1976) and certain privileges (Bailey et al., 1970; Clark, 1972; Hawkins et al., 1972, Experiment 4; Heaton et al., 1976; Schumaker et al., 1977; Strober & Bellack, 1975; Kirigin et al., Note 1). Some studies did not include sanctions at home for misbehavior at school (Hawkins et al., 1972, Experiment 4; Lahey et al., 1977). When included, sanctions usually involved loss of certain privileges such as TV, snacks, or a later bedtime (Ayllon et al., 1975; Clark, 1972; Heaton et al., 1976; Strober & Bellack, 1975). In the study by Todd, Scott, Bostow, and Alexander (1976), 3 successive "undesirable" days resulted in a 1-day suspension from school, a rather extreme punishment in our opinion.

In the study by Lahey et al. (1977), contingent parental praise at home, implemented for all children in two kindergarten classrooms, was sufficient to reduce school disruptive behavior. In contrast, Schumaker et al. (1977, Experiment 2) found that parental praise was initially effective with one disruptive adolescent but not with another. The second adolescent's school behavior did not improve until home privileges were added to parental praise. In addition, the initial improvement in the behavior of the first adolescent appeared to decline over time. A key difference in the Lahey et al. and Schumaker et al. studies is the age of the students; parental praise may be more effective the younger the student. The Schumaker et al. study is the only one that has attempted to compare different consequences (praise vs. praise plus privileges in this case) in the home. More studies are needed that examine the specific effects of praise, tangible rewards, and sanctions when used by parents as consequences for disruptive behavior at school and that examine how each of these interacts with the demographic variables (e.g., age, socioeconomic status) of the subject.

Of special importance is the wide variety of students with which some type of home-based reinforcement program has been effective in reducing disruptive behavior in the classroom. It has been used successfully with children of all ages, from elementary school to high school. The students in the studies reviewed were frequently more than 1 year behind in academic skills (Ayllon et al., 1975; Schumaker et al., 1977), had previous records of severe misconduct (Heaton et al., 1976), and were labeled as emotionally disturbed (Coleman, 1973) or underachievers (Hawkins et al., 1972, Experiment 4). Home-based reinforcement programs have even proven to be effective with delinquents in both special and regular classrooms (Bailey et al., 1970; Martin, Burkholder, Rosenthal, Tharp, & Thorne, 1968; Kirigin et al., Note 1).

Academic Behaviors in the Classroom

Academic behaviors include listening to presentations by the teacher, listening and participating in class discussions, asking and answering questions, working with eyes and head oriented toward materials, completing classwork neatly and correctly, and achieving (i.e., grades) appropriate to potential. As with disruptive behaviors, teachers have successfully used a number of behavioral techniques to improve the academic behavior of one or more students in their classroom (O'Leary & O'Leary, 1976, 1977b). Although effective, these procedures have the same limitations as those mentioned in the previous section. Not surprisingly, psychologists have successfully applied the same home-based reinforcement program that is used with disruptive behaviors to improve academic behaviors in the classroom. In fact, the use of this technique in the studies reviewed was rarely limited to either disruptive behaviors or academic behaviors. More often, some disruptive and some academic behaviors were included together in the teacher–parent communication (e.g., Bailey et al., 1970; Cohen,

Keyworth, Kleiner, & Libert, 1971; Lahey et al., 1977; Schumaker et al., 1977).

As might be expected from the large number of academic behaviors described earlier, the studies reviewed differed widely in their criteria for "good academic behavior." Some studies focused on behaviors that are considered conducive to learning (e.g., orienting eyes and head to work, looking at teacher, and responding to questions); a good teacher-to-parent report in this case indicated that the student had engaged in specific appropriate academic behaviors during class (Bailey et al., 1970; Coleman, 1973; Kirigin et al., Note 1). Other studies chose completion of classwork as the target for improvement and required the student to complete all classwork to receive a good report (Cantrell, Cantrell, Huddleston, & Wooldridge, 1969, Case 1; Harris, Finfrock, Giles, Hart, & Tsosie, 1975; Lahey et al., 1977). Several studies required not only classwork completion but also a certain level of achievement for a good report (Hawkins et al., 1972, Experiments 1–3; Stuart, 1971; Kirigin et al., Note 1; Kirigin et al., Note 2).

Although potentially effective, the use of a certain level of achievement on classwork as the criterion for good academic behavior at school and as the basis for parental consequences at home can produce certain problems. If the level of achievement is set too high, the student may be unable to meet the criterion. As a consequence, he or she has been placed in a "can't win" situation and, no matter how motivated initially, may never be able to receive any of the home rewards. Of course, a teacher–parent communication of this type would be unlikely to bring about the desired changes in the student's academic behavior.

Several studies have successfully handled the problems associated with the use of achievement level in the teacher–parent communication. In the study by Kirigin et al. (Note 2), two of the three female delinquents had no difficulty when good academic behavior and subsequent rewards at home were based on number of problems correctly done. In fact, the two girls improved their daily grade from a D to a B and a D to a C, respectively. The third female delinquent, however, showed no change in her academic behavior, even after the teacher–parent communication had been in effect for a week. When tutoring was included in the program, her accuracy on daily assignments quickly improved from an F to an A.

The home-based reinforcement program used by Hawkins et al. (1972, Experiments 1–3) provides another possible solution for using a criterion of achievement level with students performing below grade level. In their study the initial accuracy criterion for good academic behavior was placed at a level commensurate with the student's current performance. Each day the criterion was gradually adjusted upward with remarkable success.

As with disruptive behaviors, the content or degree of specificity of the teacher's report to the parents varied from global to detailed. Usually it was global, with teachers indicating only whether the student "studie[d] the whole period" (Bailey et al., 1970), "did very well" (Hawkins et al., 1972, Experiments 1–3; Heaton et al., 1976; Karraker, 1972) or "completed all his work" (Lahey et al., 1977). The study by Cantrell et al. (1969, Case 1) is an example of the other extreme. The teacher–parent communication indicated whether the student's class assignments were completed, were well-done, and had no more than two careless errors, whether the student listened to and complied with directions, and what daily grades were earned by the student. As with disruptive behaviors, there have been no studies that systematically examined the importance of degree of specificity of the teacher's report on the student's academic improvement.

In most studies that dealt with academic behaviors in the classroom, the teacher–parent communication occurred daily (e.g., Bailey et al., 1970; Karraker, 1972; Kroth et al., 1970; Schumaker et al., 1977; Kirigin et al., Note 1, Kirigin et al., Note 2). Some studies sent notes home only when the student displayed "good" academic behavior (Hawkins et al., 1972, Experiments 1–3, Heaton et al., 1976); others limited the teacher–parent contact to once a week (Coleman, 1973; Harris et al., 1975; McKenzie, Clark, Wolf, Kothera, & Benson, 1968). Again, there have been no studies that examined the frequency of teacher–parent

contact on the effectiveness of the home-based reinforcement programs.

A recent study did compare the effects of fixed-time and variable-time teacher–parent communications on the academic behavior of 26 third-grade students (Saudargas, Madsen, & Scott, 1977). Under the fixed-time condition, each student received a report for his/her parents on Friday that indicated the quantity and quality of her/his work completed that week. Under the variable-time condition, the same information was communicated to parents, but seven to nine students were randomly selected each day to receive a report to take home. The results show that more assignments were correctly completed when the students were unsure whether they would receive a report that day; that is, their academic output was higher under the variable-time condition. A similar study comparing daily to weekly teacher–parent communications is certainly needed.

The rewards and sanctions received by students at home for good academic behavior at school were similar to those in the previous section. Nearly all of the studies used home privileges (e.g., allowance, TV time, snacks, late bedtime) to reward a good teacher report (e.g., Bailey et al., 1970; Karraker, 1972; Schumaker et al., 1977). Two studies limited the rewards received at home to allowances (Coleman, 1973; McKenzie et al., 1968). This is perhaps simpler for parents to operate or at least less disruptive to family routines in that parents do not have to monitor such activities as TV time, bedtime, and so forth. Sanctions, when included, were limited to loss of privileges and were always used in combination with a reward system. As with the consequences for disruptive behaviors, more studies are needed to examine the effectiveness of different reward and punishment programs in changing academic behaviors.

Home-based reinforcement programs have been successful in increasing the academic behavior of a wide variety of students. As might be expected, the students in the studies reviewed were frequently described as under-achievers or as poorly motivated (Hawkins et al., 1972, Experiments 1–3; Karraker, 1972). Often the students were behind more than 1 year in several academic subjects (e.g., Schumaker et al., 1977). Of particular interest is the study by McKenzie et al. (1968). The students in their study had all been labeled as having learning disabilities.

Home-based reinforcement programs have also been used to increase academic behavior with students whose major difficulties may not be academic achievement. Several studies included students who had been labeled emotionally disturbed by teachers and/or mental health professionals (Coleman, 1973; Kroth, 1970). Other types of students included those considered to be moderate to severe discipline problems at school (Heaton et al., 1976; Schumaker et al., 1977; Strober & Bellack, 1975) and those considered to be delinquents by the community (Bailey et al., 1970; Harris et al., 1975; Martin et al., 1968; Stuart, 1971; Kirigin et al., Note 1, Kirigin et al., Note 2).

Methodological Evaluation

All of the studies reviewed present results that support the position that home-based reinforcement programs provide an effective treatment approach to modify both disruptive and academic behaviors in the classroom. This type of treatment has the added advantages of minimizing the time required by the teacher for implementation and involving the parents as change agents for their child's classroom behaviors. But before a complete endorsement of this treatment approach can be made, a more rigorous assessment of the research is needed to evaluate the validity of the results and to delineate areas for further investigation. To organize the critique of the studies reviewed, six categories (i.e., adequate design, systematic variation of treatment, multiple-outcome measures, follow-up, school program monitored, and home program monitored) that were deemed critical for sound research, were determined, and the articles were evaluated in terms of whether they met the requirements of each category. Four additional descriptive categories (i.e., sample size, grade, classroom setting, and target behavior) were included to assess the applicability of the results to different populations and settings. A summary of this evaluation is presented in Table 1.

Descriptive Categories

Regarding the descriptive categories, the number of subjects participating in the experiments varied from 1 to 124. The mean, median, and mode were 10.8 subjects, 2 subjects, and 1 subject, respectively. Subjects were from the primary grades (kindergarten through 6th) in 45% of the experiments and from the secondary grades (7th–12th) in the remaining 55% of the investigations. Sixty-eight percent and 32% of the experiments occurred in regular and special (e.g., remedial, emotionally disturbed, learning disability) classrooms, respectively. In 39%, 19%, and 42% of the investigations, academic (e.g., listening to presentations by teacher, asking and answering questions, working with eyes and head oriented towards materials, completing classwork, achieving, making good grades), disruptive (e.g., making noise, talking without permission, physically disturbing other children, getting out of one's seat without permission), and combined academic and disruptive behaviors, respectively, were targeted. These data indicate that most of the studies have involved few subjects; however, the programs have been implemented in all grades, in regular as well as special classrooms, and with both academic and disruptive behaviors.

Adequate Design

A major concern in evaluating the soundness of the studies reviewed is the adequacy of the experimental design that was used to examine treatment effects. The use of any of three types of designs was considered an adequate basis for the conclusion that the results were due to the experimental treatment. These were (a) ABA design (or some variation thereof with baseline and reversal), (b) multiple baseline design, and (c) group design with appropriate control group. If the experimental design of a study was one of the three above and there was not the confound of multiple interventions (e.g., home- plus school-based programs), the design was considered rigorous enough for the conclusion that the results were indeed due to the home-based reinforcement treatment program and not due to alternative factors.

Only 63% of the studies reviewed had adequate designs without multiple interventions (Table 1). Of these, the most frequently chosen design (10 of 17 or 58.8%) was some variation of the ABA design (e.g., Ayllon et al., 1975; Bailey et al., 1970). Only one study (Harris et al., 1975) selected a group design with an appropriate control group to examine the effects of home-based treatment programs. The studies with inadequate designs were usually studies with baseline and treatment data but no treatment reversal (e.g., Cantrell et al., 1969; Hawkins et al., 1972; Strober & Bellack, 1975). One study was considered inadequate because the home-based reinforcement program was confounded by being part of a larger multiple intervention treatment program. Conclusions about the effectiveness of home-based reinforcement were reached in spite of the confound (Coleman, 1973).

Systematic Variation of Treatment

A second criterion that was selected to evaluate the reviewed studies is the inclusion of systematic variation of treatment variables in the experimental design (i.e., component analysis). This is necessary if one is to differentiate which of the many variables that make up a treatment program are actually responsible for the observed changes (e.g., social vs. tangible rewards). Only when this type of information has been determined can a treatment program be refined to maximize its effectiveness while minimizing or eliminating unnecessary features.

Of the studies reviewed, 39% did include some systematic examination of the contributions of different treatment components, although their manipulations were by no means exhaustive. Two studies (Ayllon et al., 1975; Bailey, et al., 1970) investigated the effects of contingent and noncontingent reports home. In the noncontingent condition, the students received "good" reports and subsequent home-based rewards regardless of their school behavior. No treatment effects were observed until a good report was made contingent on actual school behavior. It appears that contingent consequences for target behaviors are necessary for change in

Table 1
Summary of Studies That Used Home-Based Reinforcement Programs to Modify Classroom Behaviors

Author	Experiment	N	Grade	Classroom setting	Target behavior	Adequate design	Systematic variation of treatment	Multiple-outcome measures	Follow-up	School program monitored	Home program monitored
Ayllon, Garber, & Pisor (1975)	2	23	Primary	Regular	Disruptive	Yes	Yes	No	No	Yes	No
Bailey, Wolf, & Phillips (1970)	1	5	Secondary	Special	Academic and disruptive	Yes	Yes	Yes	No	Yes	No
	2	1	Secondary	Regular	Academic	Yes	No	No	No	Yes	No
	3	1	Secondary	Regular	Adacemic and disruptive	Yes	Yes	No	No	Yes	No
Cantrell, Cantrell, Huddleston, & Wooldridge (1969)	1	1	Secondary	Regular	Academic	No	No	No	No	No	No
Clark (1972)	1	124	Primary and secondary	Special	Disruptive	No	No	No	No	No	No
	2	22	Primary and secondary	Special	Disruptive	Yes	No	No	No	Yes	No
Cohen, Keyworth, Kleiner, & Libert (1971)	1	1	Secondary	Special	Academic and disruptive	—[a]	No	Yes	No	No	No
	2	1	Secondary	Special	Academic and disruptive	—[a]	No	Yes	No	No	No
Coleman (1973)		1	Primary	Regular	Academic and disruptive	No	No	No	No	No	No
Harris, Finfrock, Giles, Hart, & Tsosie (1975)		5	Secondary	Regular	Academic	Yes	No	No	No	No	No
Hawkins, Sluyter, & Smith (1972)	1	1	Primary	Regular	Academic	No	Yes	No	No	No	No
	2	1	Primary	Regular	Academic	No	Yes	No	No	No	No
	3	1	Primary	Regular	Academic	No	Yes	No	No	No	No
	4	1	Primary	Regular	Disruptive	No	Yes	No	No	No	No
Heaton, Safer, Allen, Spinnato, & Prumo (1976)		14	Secondary	Special	Academic and disruptive	—[a]	No	Yes	No	No	No
Karraker (1972)		16	Primary	Regular	Academic	Yes	Yes	No	No	No	No
Kirigin, Bailey, Phillips, Fixen, & Wolf (Note 1)		6	Secondary	Special	Academic and disruptive	Yes	Yes	Yes	No	Yes	No

Table 1 (*continued*)

Author	Experiment	N	Grade	Classroom setting	Target behavior	Adequate design	Systematic variation of treatment	Multiple-outcome measures	Follow-up	School program monitored	Home program monitored
Kirigin et al. (Note 2)		3	Secondary	Regular	Academic	Yes	No	No	No	No	No
Kroth, Whelan, & Stables (1970)		5	Secondary	Special	Academic	Yes	No	No	No	No	No
Lahey et al. (1977)		50	Primary	Regular	Academic and disruptive	Yes	No	Yes	No	Yes	No
Martin, Burkholder, Rosenthal, Tharp, & Thorne (1968)		5	Secondary	Special	Academic and disruptive	—[a]	No	No	Yes	Yes	No
McKenzie, Clark, Wolf, Kothera, & Benson (1968)		10	Primary	Special	Academic	No	No	No	Yes	Yes	No
Saudargas, Madsen, & Scott (1977)		26	Primary	Regular	Academic	Yes	Yes	No	No	Yes	No
Schumaker, Hovell, & Sherman (1977)	1	3	Secondary	Regular	Academic and disruptive	Yes	No	Yes	No	No	No
	2	2	Secondary	Regular	Academic and disruptive	Yes	Yes	Yes	No	No	No
	3	2	Secondary	Regular	Academic and disruptive	Yes	No	Yes	No	No	No
Strober & Bellack (1975)		1	Primary	Regular	Academic and disruptive	No	No	No	Yes	No	No
Stuart (1971)		1	Secondary	Regular	Academic	No	No	No	No	No	No
Todd, Scott, Bostow, & Alexander (1976)[b]	1	1	Primary	Regular	Disruptive	Yes	No	No	Yes	Yes	No
	2	1	Primary	Regular	Disruptive	Yes	Yes	No	Yes	Yes	No
Criterion met[c]						63.0%	38.7%	29.0%	16.1%	38.7%	0%

[a] This study was confounded by multiple interventions. However, the primary purpose of the study was not to examine the effects of home-based reinforcement; therefore, the study was not considered in the adequate design category.
[b] Some of information presented for the Todd et al. study is based on Bostow (Note 3).
[c] Percentage of the number of experiments that met the criterion in each methodological category.

those behaviors. Two other studies (Hawkins et al., 1972; Karraker, 1972) examined a somewhat similar variable, feedback versus feedback plus home-based consequences. Teacher feedback to the child and his/her parents concerning the child's school behavior was not effective unless the feedback was linked with consequences at home. Other treatment components that have been studied are schedule of report home (Saudargas et al., 1977), types of home-based consequences (Schumaker et al., 1977), and types of behaviors targeted for treatment (Kirigin et al., Note 1).

Multiple-Outcome Measures

Another criterion deemed necessary to evaluate treatment effectiveness is the inclusion of multiple-outcome measures. Different outcome measures frequently lead to different conclusions (Forehand & Atkeson, 1977). To provide a complete evaluation of the effectiveness of a home-based program, one should not only measure changes in a child's behavior (e.g., attending to teacher's presentation, talking without permission) and his or her output (e.g., assignment completion, grades) but should also measure changes in the teacher's and parents' perceptions of the child. Therefore, dependent measures that adequately assess the multiple outcomes of treatment (e.g., child behavior change, adult perception change) are required.

Only 29% of the studies reviewed included more than one assessment measure to evaluate their treatment program. Of these studies, only two evaluated the program's effect on the teacher's (Schumaker et al., 1977) or the parents' (Lahey et al., 1977) perception of the child. The remaining studies that used multiple-outcome measures limited their assessment of treatment effect to two measures of the child: (a) changes in the child's classroom behavior and (b) changes in her or his academic output.

Follow-Up

The fourth methodological criterion is the inclusion of follow-up measures to determine the temporal generality of home-based reinforcement programs on school behavior. Even if a treatment program can produce immediate changes in the targeted behaviors, it is of little benefit if these effects are transitory or if they reverse once treatment is terminated. Only 16% of the studies reviewed included follow-up measures to evaluate the long-term effectiveness of their programs. When included, these measures were frequently not rigorous enough to adequately assess temporal generality. For example, some studies examined treatment effects following termination but presented no data (e.g., Strober & Bellack, 1975), and when follow-up data were presented, the measures that were used often were not the same as those used during treatment (e.g., McKenzie et al., 1968). More data are needed before we conclude that home-based reinforcement programs produce lasting effects. In addition, treatment variables that might enhance temporal generality (e.g., fading of treatment) need to be systematically examined.

School Program Monitored and Home Program Monitored

The last two methodological criteria involve treatment implementation. A prerequisite for the evaluation of the effectiveness of treatment is the determination of whether the program was in fact implemented as planned. Home-based reinforcement for school behavior involves treatment implementation in both the school and home; therefore, the behaviors of the change agents in both these settings should be monitored to determine if they are conforming to the program outlined. This evaluation can take several forms. In some studies, the researchers maintained telephone contact with the parents (e.g., Ayllon et al., 1975), others had periodic conferences (e.g., Cohen et al., 1971; Heaton et al., 1976; Kroth et al., 1970; McKenzie et al., 1968), and several placed trained observers in the treatment setting (e.g., Bailey et al., 1970; Lahey et al., 1977; Saudargas et al., 1977). For purposes of methodological evaluation, we selected the last, stricter criterion (i.e., use of observers) as necessary to monitor treatment implementation adequately.

In the school setting 39% of the studies reviewed had observers present. In contrast, none of the studies reviewed monitored treatment implementation in the home with observational data. This lack of control makes it difficult to infer conclusively that the behavioral changes observed in the classroom were in fact due to the treatment and not to other factors in the home. Although this omission in experimental design probably arises from the many difficulties associated with placing trained observers in the home, researchers cannot ignore the necessity for some type of assessment of program implementation in this setting. One solution to this dilemma is currently being examined by Bostow (Note 3). In his study, parents are instructed to audiotape the parent–child interaction when the child receives her/his home consequences that are based on the teacher communication.

Conclusions

The need to involve parents in the school-related problems of their children is evident. Parents are reported to be experiencing a loss of influence over their children as schools are assuming roles once reserved for parents (Woodward, 1978). Home-based reinforcement offers one way in which parents can be incorporated into programs designed to manage the school difficulties of their children. The procedure permits parents to receive regular feedback concerning their child's school behavior. Furthermore, home-based reinforcement encourages frequent teacher–parent communication. From the school's perspective, home-based reinforcement eliminates many of the ethical issues associated with school behavior modification programs, since neither disciplinary procedures nor tangible positive reinforcement have to be used in the school. Furthermore, the procedure circumvents the time-consuming and difficult process of setting up a behavioral classroom program.

In all of the studies reviewed, the conclusion reached by the respective authors was that home-based reinforcement was effective in changing classroom behavior. It is encouraging to note that this conclusion was consistent across a wide range of grades, regular and special classrooms, and academic and disruptive behaviors. The replication of the effectiveness of home-based reinforcement across this variety of ages, settings, and behaviors attests to the general impact of the procedure.

Unfortunately, an analysis of the methodology used in the studies that examined home-based reinforcement yields a less positive picture. Appropriate designs have been used in less than two-thirds of the studies. Furthermore, monitoring of the school and home programs has rarely occurred. Behavior modifiers typically have prided themselves in being well grounded in experimental methodology. Indeed, articles and books have been devoted to the topic (e.g., Birnbrauer, Peterson, & Solnick, 1974; Hersen & Barlow, 1976). Unfortunately, in the case of home-based reinforcement programs, the methodology often has been less than adequate. Other aspects of assessment that are receiving increasing emphasis in behavior modification, such as multiple-outcome measures (e.g., Atkeson, & Forehand, 1978; Turkat & Forehand, in press) and follow-up measures (e.g., Forehand & Atkeson, 1977), have been ignored in the home-based reinforcement studies. Obviously, unless adequate methodologies are employed, our conclusions about the effectiveness of home-based reinforcement procedures will be limited.

Reference Notes

1. Kirigin, K. A., Bailey, J. S., Phillips, E. L., Fixen, D. L., & Wolf, M. M. *The effects of home-based reinforcement on the study behavior and academic performance of pre-delinquent boys.* Unpublished manuscript, University of Kansas, 1971.
2. Kirigin, K. A., et al. *The effects of home-based reinforcement on the modification of academic behavior of three delinquent girls.* Paper presented at the meeting of the American Psychological Association, Montreal, Canada, August 1973.
3. Bostow, D. E. Personal communication, April 20, 1978.

References

Atkeson, B. M., & Forehand, R. Parent behavioral training: An examination of studies using multiple outcome measures. *Journal of Abnormal Child Psychology*, 1978, *6*, 449–460.

Ayllon, T., Garber, S., & Pisor, K. The elimination of discipline problems through a combined school-home motivational system. *Behavior Therapy*, 1975, *6*, 616–626.

Bailey, J. S., Wolf, M. M., & Phillips, E. L. Home-based reinforcement and the modification of pre-delinquents' classroom behavior. *Journal of Applied Behavior Analysis*, 1970, *3*, 223–233.

Berkowitz, B. P., & Graziano, A. M. Training parents as behavior therapists: A review. *Behaviour Research and Therapy*, 1972, *10*, 297–317.

Birnbrauer, J. S., Peterson, C. R., & Solnick, J. V. Design and interpretation of studies using single subjects. *American Journal of Mental Deficiency*, 1974, *79*, 191–203.

Cantrell R. P., Cantrell, M. L., Huddleston, C. M., & Wooldridge, R. L. Contingency contracting with school problems. *Journal of Applied Behavior Analysis*, 1969, *2*, 215–220.

Clark, H. B. A program of delayed consequences for the management of class attendance and disruptive classroom behavior of 124 special education children. In G. Semb (Ed.), *Behavior analysis and education*. Lawrence: University of Kansas, 1972.

Cohen, S., Keyworth, J., Kleiner, R., & Libert, J. The support of school behaviors by home-based reinforcement via parent-child contingency contracts. In E. Ramp & B. Hopkins (Eds.), *A new direction for education: Behavioral analysis*. Lawrence: University of Kansas, 1971.

Coleman, R. G. A procedure for fading from experimenter-school-based to parent-home-based control of classroom behavior. *Journal of School Psychology*, 1973, *11*, 71–79.

Forehand, R., & Atkeson, B. M. Generality of treatment effects with parents as therapists: A review of assessment and implementation procedures. *Behavior Therapy*, 1977, *8*, 575–593.

Graziano, A. M. Parents as behavior therapists. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 4). New York: Academic Press, 1977.

Harris, V. W., Finfrock, S. R., Giles, D. K., Hart, B. M., & Tsosie, P. C. The effects of performance contingencies on the assignment completion behavior of severely delinquent youth. In. E. A. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research and application*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Hawkins, R. P., Sluyter, D. J., & Smith, C. D. Modification of achievement by a simple technique involving parents and teacher. In M. B. Harris (Ed.), *Classroom uses of behavior modification*. Columbus, Ohio: Charles E. Merrill, 1972.

Heaton, R. C., Safer, D. J., Allen, R. P., Spinnato, N. C., & Prumo, F. M. A motivational environment for behaviorally deviant junior high school students. *Journal of Abnormal Child Psychology*, 1976, *4*, 263–275.

Hersen, M., & Barlow, D. H. *Single case experimental designs*. New York: Pergamon Press, 1976.

Johnson, C. A., & Katz, R. C. Using parents as change agents for their children: A review. *Journal of Child Psychology and Psychiatry*, 1973, *14*, 181–200.

Karraker, R. J. Increasing academic performance through home-managed contingency programs. *Journal of School Psychology*, 1972, *10*, 173–179.

Kroth, R. L., Whelan, R. J., & Stables, J. M. Teacher application of behavior principles in home and classroom environments. *Focus on Exceptional Children*, 1970, *2*, 1–10.

Lahey, B. B., et al. An evaluation of daily report cards with minimal teacher and parent contacts as an efficient method of classroom intervention. *Behavior Modification*, 1977, *1*, 381–394.

Martin, M., Burkholder, R., Rosenthal, T. L., Tharp, R. G., & Thorne, G. L. Programming behavior change and reintegration into school milieux of extreme adolescent deviates. *Behaviour Research and Therapy*, 1968, *6*, 371–383.

McKenzie, H., Clark, M., Wolf, M., Kothera, R., & Benson, C. Behavior modification of children with learning disabilities using grades as tokens and allowances as back-up reinforcers. *Exceptional Children*, 1968, *43*, 745–752.

O'Dell, S. Training parents in behavior modification: A review. *Psychological Bulletin*, 1974, *81*, 418–433.

O'Leary, K. D., & O'Leary, S. G. Behavior modification with children. In K. D. O'Leary & S. G. O'Leary (Eds.), *Classroom management: The successful use of behavior modification*. New York: Pergamon Press, 1977. (a)

O'Leary, K. D., & O'Leary, S. G. (Eds.). *Classroom management: The successful use of behavior modification*. New York: Pergamon Press, 1977. (b)

O'Leary, S. G., & O'Leary, K. D. Behavior modification in the school. In H. Leitenberg (Ed)., *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.

Saudargas, R. W. Madsen, C. H., Jr., & Scott, J. W. Differential effects of fixed- and variable-time feedback on production rates of elementary school children. *Journal of Applied Behavior Analysis*, 1977, *10*, 673–678.

Schumaker, J. B., Hovell, M. F., & Sherman, J. A. An analysis of daily report cards and parent-managed privileges in the improvement of adolescents' classroom performance. *Journal of Applied Behavior Analysis*, 1977, *10*, 449–464.

Strober, M., & Bellack, A. S. Multiple component behavioral treatment for a child with behavior problems. *Journal of Behavior Therapy and Experimental Psychiatry*, 1975, *6*, 250–252.

Stuart, R. B. Behavioral contracting within the families of delinquents. *Journal of Behavior Therapy and Experimental Psychiatry*, 1971, *2*, 1–11.

Todd, D. D., Scott, R. B., Bostow, D. E., & Alexander, S. B. Modification of the excessive inappropriate classroom behavior of two elementary school students using home-based consequences and daily report-card procedures. *Journal of Applied Behavior Analysis*, 1976, *9*, 106.

Turkat, I. D., & Forehand, R. Critical issues in behavior therapy. *Behavior Modification*, in press.

Woodward, K. L. Saving the family. *Newsweek*, May 15, 1978, pp. 63–73.

Received July 3, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1309–1321

Sex Roles and Psychotherapy: A Current Appraisal

Bernard E. Whitley, Jr.
University of Pittsburgh

This article reviews the current status of research on the effects of sex role stereotypes on mental health judgments. Studies in this area have addressed three questions: (a) Are there different, sex-role-related standards of mental health for men and women? (b) Do violations of sex role norms result in adverse mental health judgments? (c) Do therapists set sex-role-related goals for their clients? It is concluded that sex role stereotypes are strong mental health cues for nonprofessionals, with violations of sex role norms leading to adverse mental health judgments, but that whereas professionals share the sex role stereotypes of their lay contemporaries, the professionals are unaffected by them in making mental health judgments and in setting therapeutic goals. This discrepancy between stereotypes and behavior may be due to any of three factors: the methodological limitations of the studies, actual differences in mental health between men and women, or normal attitude–behavior discrepancies.

One of the more popular villains in contemporary psychology is the psychotherapist as a covert (if unwitting) agent of social control and the status quo (e.g., Hurvitz, 1973; Leifer, 1970; Szasz, 1961). This thesis has been most vigorously expounded by a number of feminist writers (e.g., American Psychological Association, 1975; Chesler, 1972; De Beauvoir, 1949/1967; Tennov, 1975) who have charged that therapists define mental health in terms of sex role stereotypes and impose those stereotypes on their clients under the guise of therapy, thereby inhibiting rather than facilitating mental health. If true, these charges are indeed serious, for research has shown that overadherence to stereotyped sex roles is associated with psychopathology (e.g., H. Goldberg, 1976; Gove, 1972; Gove & Tudor, 1973) and low self-esteem (e.g., Bem, 1977; Spence, Helmreich, & Stapp, 1975).

These allegations of sex role bias were supported only by intuition and case histories until the publication of a study by Broverman, Broverman, Clarkson, Rosenkrantz, and Vogel (1970). Broverman and her collegues found that psychotherapists shared the sex role stereotypes of their society and that they included those stereotypes as part of their definition of mental health, thus providing some evidence of therapeutic bias. They did not, however, investigate the extent to which this attitudinal bias resulted in discriminatory behavior. Subsequent investigation of such bias and behavior has generally found no such discrimination and casts doubt on the generality of the bias. This article reviews that subsequent research.

Questions

Within this research, three questions have been of special importance: (a) Are there different, sex-role-related standards of mental health for men and women? (b) Do violations of sex role norms result in adverse mental health judgments? (c) Do therapists set sex-role-related goals for their clients? Although this article reviews the current status of research on these questions, it should be noted that this review only considers research on sex roles in mental health judgments, not on simple sex-of-client effects. Research on the latter's relationship to mental health judgments, test-

---

The author is grateful to Irene Hanson Frieze, Paul A. Pilkonis, and Walter R. Gove for discussion of earlier drafts of this manuscript.

Requests for reprints should be sent to Bernard E. Whitley, Jr., Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

 0033-2909/79/8606-1309$00.75

Table 1
Studies That Investigated Differential Mental Health Standards

Study	Instrument	Findings	Remarks
Mental health professionals as subjects			
Broverman, Broverman, Clarkson, Rosenkrantz, & Vogel (1970)	BSRQ	+	Dichotomous scoring
Fabrikant, Landau, & Rollenhagan (1973)	ACL	+	
	FR	−	
Fabrikant (1974)	ACL	+	
	FR	−	
Johnson (1974)	BSRQ	+	
Anderson (1975)	BSRQ	+	
Maxfield (1976)	BSRQ	+/−	+ for dichotomous scoring only
Aslin (1977)	BSRQ	+/−	+ for males rating females only
Mental health trainees as subjects			
Terrill (1972)	BSRQ	+	
Maslin & Davis (1975)	BSRQ	−/+	+ for males rating females only
Harris & Lucas (1976)	BSRQ	−/+	+ for males rating females only
Other subject groups			
Nowacki & Poe (1973)	BSRQ	+	College students
Kravetz (1976)	BSRQ	−	College women only

Note. BSRQ = Broverman et al. (1970) sex role questionnaire; ACL = Adjective Check List; FR = free response; + = supports hypothesis; − = does not support hypothesis.

ing, and treatment has been briefly reviewed by Abramowitz and Dokecki (1977) and Zeldow (1978).

Subject Populations

Researchers have used three populations of subjects in studying the relationship of sex role stereotypes to mental health judgments: mental health professionals, professional trainees, and nonprofessionals. The first two categories typically include psychiatrists, clinical psychologists, counselors, and social workers. Although most research has been directed at the first two population groups, this review includes research on nonprofessionals as well, both for comparative purposes and because they influence the client referral process (cf. Scott, 1958).

Stereotypes As Standards of Mental Health

Twelve studies have tested the hypothesis that there are different standards of mental health for men and women, which are based on sex role stereotypes (Table 1).[1] Subjects in these studies indicated the chracteristics of the mentally healthy man and woman in one of three ways: by giving a free response, by marking a trait checklist, or by rating the degree to which the stimulus persons possess certain traits (continuous scale). Generally, these studies have tended to support the hypothesis, but many contain flaws that limit their validity.

Results for Different Subject Populations

Mental Health Professionals

The Broverman et al. study. The first, and most cited, study linking sex role stereotypes to mental health standards was that conducted

[1] Not included in this review is a study by Shapiro (1977) that used subjects from a population that the author described as potentially biased.

by Broverman et al. (1970). The psychiatrists, psychologists, and social workers who participated in the study responded to a modified version of a sex role stereotype questionnaire developed by Rosenkrantz, Vogel, Bee, Broverman, and Broverman (1968). The modified questionnaire consisted of 122 bipolar adjective pairs, of which 27 had been determined to be stereotypically masculine and 11 to be stereotypically feminine. Subjects completed the questionnaire by indicating "the pole to which a mature, healthy, socially competent" adult, man, or woman would be closer (Broverman et al., 1970, p. 2). Masculine and feminine health scores were computed relative to the traits assigned to the adult, and significant differences were found between the mean male and female health scores.

Replications of the Broverman et al. study. Fabrikant (1974) and his collegues (Fabrikant, Landau, & Rollenhagen, 1973) attempted to replicate the Broverman et al. (1970) study using an independently developed checklist and free-response items. These studies found sex-role-related mental health standards on the checklist but not on the free-response items.

The results of studies that used continuous scales are mixed. Although Anderson (1975) and Johnson (1974) found stereotyped mental health standards using the Broverman et al. (1970) sex role questionnaire (BSRQ), as did Aslin (1977) for male therapists rating women, Maxfield (1976) did not.

Professional Trainees

Attempts to identify stereotypic mental health conceptualizations among professional trainees have also met with mixed results. Terrill (1972) elicited such stereotypes using a continuous-scale BSRQ with counselor trainees, but Maslin and Davis (1975) and Harris and Lucas (1976) found few such stereotypes with counselor and social work trainees, respectively. Their only statistically significant difference was between male conceptualizations of females and the other subject-stimulus combinations.

Other Subject Populations

Nowacki and Poe (1973) investigated the mental health concepts of college students using a continuous-scale BSRQ. Differences were found for the concepts of mentally healthy man and woman; differences among male subjects were more extreme. Kravetz (1976) employed a continuous-scale BSRQ with a sample of college women. No differences were found between the concepts of mentally healthy man and woman on 73% of the "stereotypic" items, nor were there overall differences. Self-reported membership of some of the subjects in the women's liberation movement had no effect on the results.

Moderating Variables

Variables that have been postulated to interact with the sex of the target to produce stereotyping of mental health concepts include the judge's sex, personal style, and role (therapist vs. nontherapist).

Sex of Judge

As noted above, several studies have found sex of subject effects in conceptualizations of mental health (Aslin, 1977; Harris & Lucas, 1976; Maslin & Davis, 1975; Nowacki & Poe, 1973; also Delk & Ryan, 1977). The usual result is that men tend to stereotype to a greater degree than do women, particularly when men rate women.

Personal Style

A–B therapist status. Delk and Ryan (1975, 1977) have also studied the effects of A–B status on stereotyping. Among therapists, As are more successful at treating schizophrenics, Bs are more successful at treating neurotics, and As tend to attribute more feminine characteristics to themselves than do Bs (Delk & Ryan, 1977). The studies found that As also tend to stereotype more than do Bs.

Personal stereotypes. Comparing therapists' personal sex role stereotypes with the cultural stereotypes, Billingsley (Note 1) found four types of therapists: (a) those whose personal stereotypes accurately reflected the cultural stereotype but included a large number of other beliefs, (b) those whose personal stereotypes were largely inaccurate compared to the cultural stereotypes and included a moderate

number of other items, (c) those who had a moderately accurate personal stereotype and included a moderate number of other items, and (d) those who had an inaccurate personal stereotype with few outside items. Of the four groups, the third differentiated more between mentally healthy men and women, as measured by a continuous-scale BSRQ, than did the other three groups. However, the significance of the differences was not reported.

Subject Population

Delk and Ryan (1977) found that mental patients tended to differentiate between the traits assigned to mentally healthy men and women more than did students, who did so to a greater extent than did therapists. The difference between the patient group and the others may have been due to education as well as patient status, since they were mostly from the lower and lower-middle socioeconomic classes.

Methodological Limitations

Although these studies generally support the hypothesis that there are different standards of mental health for men and women, they have a number of methodological flaws that limit their validity. These limitations can be classified as those concerned with the measurement of stereotypes and those concerned with the scoring of the stereotype measures.

Measurement of Sex Role Stereotypes

The goal of stereotype measurement is to determine the "set of beliefs a person holds about most members of a particular social group" (Oskamp, 1977, p. 124). This task presents two problems: whether the beliefs should be assessed with a free-response or forced-choice format and whether the important beliefs are those about specific behaviors or more generalized personality traits.

Free response versus forced choice. In the free-response situation, subjects describe the mentally healthy man or woman using their own words, whereas in the forced-choice situation, they describe the mentally healthy man or woman using a list of traits provided for them, some of which are sex role stereotypic. The forced-choice response format maximizes stereotypic responses, since the subjects are allowed to respond only with the items listed on the questionnaire, regardless of whether they would ordinarily use them (cf. Lunneborg, 1970). Frieze (Note 2), for example, found that college men and women both gave primarily nonstereotypic responses to stimuli such as "I believe that most women . . .." This phenomenon was also described by Fabrikant (1974; Fabrikant et al., 1973), who found less stereotyping by therapists with free-response items than with an adjective checklist.

Behaviors and traits. The phrasing of the items on a stereotype questionnaire as either specific behaviors or general traits can affect stereotype measurements. Komarovsky (1976) and Steinman and Fox (1966) found that male college students expressed generally nonstereotypic views of women on trait-rating scales but gave more traditional responses when questioned about specific aspects of the female role, such as marriage, motherhood, and work outside the home. Thus the use of trait scales could reduce measured stereotyping.

Use of the BSRQ. Overreliance on the BSRQ itself is a major limitation of the studies described. As a forced-choice measure, the BSRQ has aspects that both increase (the forced-choice format) and decrease (use of traits) stereotyping. Without direct comparison with other instruments, the interpretation of results from the BSRQ is problematical.

Although the BSRQ was used in 10 of the 12 studies reviewed, it hardly represents a consensual definition of sex roles. Sex role inventories have also been developed by Bem (1974) and Spence et al. (1975), and the three questionnaires contain a total of 42 feminine and 53 masculine traits, of which only 2 feminine and 7 masculine traits are common to all three. In addition, only the Bem questionnaire has its masculine and feminine items matched for social desirability and has been validated against behavioral criteria (Bem, 1974, 1977). The results of a study that uses the Bem (1974), Spence et al. (1975), or a combined questionnaire could be different from one using the BSRQ.

Scoring the Stereotype Measures

Forced-choice questionnaires can be scored on either a dichotomous or a continuous scale. Using a dichotomous scale, subjects can report whether they think a trait is characteristic of a man or a woman but not the degree to which it is characteristic. This scoring system was used in the Broverman et al. (1970) and adjective checklist portions of the Fabrikant (1974; Fabrikant et al., 1973) studies. This procedure tends to inflate the extremity of scores by eliminating the possibility of a qualified or neutral rating. Investigating the implications of a dichotomous versus a continuous scale, Maxfield (1976) found a marked reduction in the rating of male–female differences when they were measured on a continuous-scale version of the BSRQ, with no statistical difference on 59% of the stereotypic items. It would thus appear that the Broverman et al. (1970) and Fabrikant (1974; Fabrikant et al., 1973) checklist findings should be interpreted with a great deal of caution.

Even when continuous scales are used, there are problems of interpretation, since most sex differences found on individual items of mental health questionnaires are of degree rather than kind (e.g., Johnson, 1974; Kravetz, 1976; Maslin & Davis, 1975; Maxfield, 1976). That is, given a bipolar adjective continuum, men and women are usually rated on the same side of the neutral point. Although the pictures of the mentally healthy man and woman may differ statistically under these conditions, the conceptual differences to the respondant may not be so great.

Summary

Although the findings of the studies that tested the hypothesis that there are different, sex-role-related standards of mental health for men and women are generally positive, methodological shortcomings cast doubt on their validity. At best, it çan be said that when equivalent measures are used, mental health professionals share the sex role stereotypes of their lay contemporaries.

Effect of Violations of Sex Role Norms on Judgments of Mental Health

If the hypothesis that there are different, sex-role-related standards of mental health for men and women is accepted, then a second hypothesis follows: The exhibition of cross-sex-role behavior leads to adverse judgments of mental health. This hypothesis has been tested in the 24 studies listed in Table 2. Research in this area has been of two types: analogue studies, in which subjects base their judgments on fictitious cases specially constructed to coincide with the independent variables of the study, and field studies, in which actual therapist–client relationships are examined. Generally, the hypothesis has been supported for nonprofessionals but not for mental health professionals.

Analogue Studies

In analogue studies, subjects are typically presented with a written description of a stimulus person whose sex role orientation is varied between judges. These stimulus persons are then rated on adjustment, mental health, or similar measures. There are three major subcategories of analogue studies: those that use only female stimulus persons, those that use both male and female stimulus persons and in which only one or two characteristics are varied to manipulate sex role orientation, and those that use both male and female stimulu s persons and in which several characteristics are varied.

Mental Health Professionals As Subjects

Female stimulus persons only. The four studies that used only female stimulus persons manipulated sex role orientation by varying one trait: Thomas and Stewart (1971), Abramowitz et al. (1975), and Hill, Tanney, Leonard, and Reiss (1977) used career choice as the critical trait, whereas Abramowitz, Abramowitz, Jackson, and Gomes (1973) used political activism. The results of these studies were generally negative, with the exception that Abramowitz et al. (1973; Abramowitz et al., 1975) found that subjects with more traditional attitudes tended to rate nontraditional stimulus persons lower on adjustment than did less traditional subjects.

Male and female stimulus persons, single trait. In these studies, sex role orientation of the stimulus persons was manipulated by varying the traits of independence and achievement (Pringle, 1973) and active versus passive

Table 2
Studies That Investigated Mental Health Judgments

Study	Findings	Remarks
	Analogue studies	
	Mental health professionals as subjects	
Thomas & Stewart (1971)	–	
Abramowitz, Abramowitz, Jackson, & Gomes (1973)	–/+	+ for subjects with traditional attitudes
Bilick (1973)	–	
Pringle (1973)	–	
Abramowitz et al. (1975)	–/+	+ for subjects with traditional attitudes
Chasen (1975)	–	
Chasen & Weinberg (1975)	–	
Berland (1976)	–	
Fischer, Dulaney, Fazio, Hudak, & Zinotofsky (1976)	–	
Gomes & Abramowitz (1976)	–	
Maxfield (1976)	–	
Hill, Tanney, Leonard, & Reiss (1977)	–	
Tribich (1977)	–/+	+ for males responding to a crisis in a feminine manner
	Mental health trainees as subjects	
Feinblatt & Gold (1976)	–/+/+	– for diagnosis, + for two prognoses
	Other subject groups	
Burhenne (1972)	+	College students
Coie, Pennington, & Buckley (1974)	+	College students
Costrich, Feinstein, Kidder, Marecek, & Pascale (1975)	+/+	College students; two experiments
Derlega & Chaikin (1976)	+	College students
Feinblatt & Gold (1976)	+	Parents
Zeldow (1976)	–/+	College students; + for males rating females only
Tribich (1977)	–/+	Adults, + for males responding to a crisis in a feminine manner
Israel, Raskin, Libow, & Pravder (1978)	+	College students
	Field studies	
	Referrals	
Collins & Sedlacek (1974)	+	Archival
Feinblatt & Gold (1976)	+	Archival
	Therapy	
Levy & Doyle (1974)	+	Interviews
Cowan (1976)	–	Survey

Note. + = supports hypothesis; – = does not support hypothesis.

personal style (Chasen, 1975; Chasen & Weinberg, 1975; Fischer, Dulaney, Fazio, Hudak, & Zinotofsky, 1976). None of these studies supported the hypothesis that cross-sex-role behavior leads to adverse judgments of mental health.

Male and female stimulus persons, multiple traits. Once again, the hypothesis was generally not supported (Berland, 1976; Bilick, 1973; Gomes & Abramowitz, 1976; Maxfield, 1976). However, Tribich (1977) found that males who responded to a crisis with crying and other "feminine" reactions were rated as more disturbed than a woman who reacted in the same manner and as more disturbed than either a man or a woman responding with anger and other "masculine" behaviors. In addition, Feinblatt and Gold (1976) found that although graduate students in clinical and school psychology did not rate the severity of a

problem greater in cross-sex-role children, they did make a stronger recommendation for treatment and a bleaker prediction for future adjustment if the behavior continued.

Other Subject Populations

Male and female stimulus persons, single trait. Studies that used college students as subjects found evidence in support of the hypothesis. In two experiments that used the traits active and passive, Costrich, Feinstein, Kidder, Marecek, and Pascale (1975) found that cross-sex-role stimulus persons were rated as being in greater need of therapy than their in-role counterparts, and Derlega and Chaikin (1976) found lower adjustment ratings for violators of self-disclosure norms. In a complex study of variables that affect mental health judgments, Coie, Pennington, and Buckley (1974) found marginally greater ratings of mental disturbance for female stimulus persons who reacted aggressively to stressful situations and for male stimulus persons who reacted to the same situations with somatic complaints.

Male and female stimulus persons, multiple traits. Using college students as subjects, Burhenne (1972) and Israel, Raskin, Libow, and Pravder (1978) found lower mental health ratings for cross-sex-role stimulus persons. Zeldow (1976), however, found sex role effects only for male subjects rating female stimulus persons. Feinblatt and Gold (1976) had a sample of parents rate the same cases as did the graduate students in another part of their study and found ratings of greater problem severity and less future adjustment for cross-sex-role children. Tribich (1977) also had nonprofessional adults rate the same cases as did his therapist subjects and again found that feminine behaviors in men led to more severe mental health ratings.

Moderating Variables

Sex of judge. Generally speaking, sex of rater has not affected judgments of mental health in relation to violations of sex role norms (Bilick, 1973; Chasen, 1975; Chasen & Weinberg, 1975; Derlega & Chaikin, 1976; Feinblatt & Gold, 1976; Gomes & Abramowitz, 1976; Maxfield, 1976; Tribich, 1977). However, woman therapists tend to be slightly more accepting of counterstereotypic behavior than men (Chasen, 1975; Chasen & Weinberg, 1975) and more lenient in their ratings of all stimulus persons (Abramowitz et al., 1975; Harris & Lucas, 1976; Maxfield, 1976). Israel et al. (1978), however, found female college students to make more severe ratings than male students.

Attitudes. Abramowitz and his collegues (1973; Abramowitz et al., 1975), working with female stimulus persons only, found that counselors with more traditional attitudes tended to attribute less adjustment to women with deviant career goals and political attitudes. Other studies (Chasen, 1975; Gomes & Abramowitz, 1976), however, have failed to replicate the attitude–sex role interaction for stimulus persons with deviant and conforming sex role traits. Although these differences may be due to differences in population (counselors vs. psychologists), the differences in stimulus material are also striking: The counselors in the first studies rated behaviors, whereas the psychologists in the attempted replications rated traits.

Situations. Although Coie et al. (1974) found marginal support for the use of sex role stereotypes in mental health judgments made by college students, they also found that the situations that elicited the judged behaviors and the situation–behavior interaction were more important cues than the stereotyping of the reaction. For example,

> relatively little disorder was attributed for social withdrawal compared to aggression in the context of exam pressure, whereas social withdrawal was seen as evidencing at least as much disorder as aggression in the context of rejection. (p. 563)

This finding suggests that situational context as well as behavioral description may be an important factor in the experimental materials.

Field Studies

Archival Studies

Two of the field studies of sex roles and judgments of mental health are archival. That is, they examined the records of a child guidance clinic (Feinblatt & Gold, 1976) and a university counseling center (Collins &

Sedlacek, 1974) for patterns of referrals related to sex role behavior. It should be noted that these studies examined requests for treatment (made directly by college students or by parents for children), not acceptance for treatment, and thus they reflect nonprofessional judgments of mental health. Collins and Sedlacek's analysis of screening interviews found that male students tended to seek counseling for vocational, educational, and underachievement problems, whereas female students sought counseling for emotional and social conflicts. Feinblatt and Gold found that 33% of the children seen at a guidance center were referred for sex-role-related behaviors and that 89% of these were for cross-sex-role behaviors. Thus failure to meet behavioral sex role expectations may be seen as a mental health problem by nonprofessionals.

Surveys

Cowan (1976) surveyed a sample of primarily male consulting psychologists by having them complete the BSRQ for the typical male and female patient, "indicating the extent to which they think one of the two poles represents the greater problem" (p. 120). She found that female clients were rated as being too feminine—less than optimum on masculine traits and more than the optimum on feminine traits—whereas male clients were not viewed in sex role terms as measured by the BSRQ. A measurement qualification arises from the fact that the psychologists, in response to another question, reported that sex role expectations underlay the problems of their male and female clients to the same extent, a finding that may reflect the inadequacy of the BSRQ as a definition of sex roles.

Interviews

On the other hand, Levy and Doyle (1974) found a slight tendancy for staff members at a drug abuse rehabilitation center to see residents' problems in sex role terms. These data, however, were scored dichotomously and reflect the views of nonprofessionals as well as professional staff, both conditions tending to increase the degree of stereotyping found.

Methodological Limitations

The weight of experimental evidence tends to disconfirm the hypothesis that violations of sex role norms result in adverse mental health judgments made by professionals and to support it for judgments by nonprofessional judges. There are, however, aspects of the stimulus material and dependent variables in the analogue studies that merit a closer look.

Independent Variable Manipulation

Specific sex-role-deviant behaviors seem to lead to more adverse judgments than trait lists. For example, Abramowitz et al. (1973; Abramowitz et al., 1975) found an interaction between counselor attitudes and stimulus behavior that led to lower adjustment ratings, whereas studies that used traits did not (Chasen, 1975; Gomes & Abramowitz, 1976). With nonprofessional subjects, Derlega and Chaikin (1976) found a much stronger effect using a specific behavior than did the other investigators who used traits. In cases such as these, behavioral description may add realism and dimension to the experimental situation. For example, the trait *very aggressive* is vague, covering a wide variety of behaviors, whereas the description "Joe pushed Tom down a flight of stairs" gives a more precise measure of aggression and allows the rater to judge the appropriateness of the response to the situation (cf. Coie et al., 1974). Degree and appropriateness of behavior are probably more salient mental health cues than traits and should be used more often in mental health studies.

Dependent Measures

Feinblatt and Gold (1976) found that clinical and school psychology graduate students rated children showing in- and cross-sex-role behavior about neutral in respect to severity of problem but that they recommended treatment more urgently and predicted less future adjustment for the cross-sex-role children. It would thus appear that ratings of present mental health, need for treatment, and future mental health elicit different judgments. It is reasonable to decide that a person is relatively healthy now but may get worse

and so needs treatment. This situation suggests the need for multivariate dependent measures of mental health that assess present perceptions and future expectations.

Summary

Mental health professionals are relatively uninfluenced by violations of sex role expectations in making mental health judgments. On the other hand, nonprofessional judges are affected by such violations when making similar decisions. These therapist–lay differences may be due to the fact that therapists stereotype to a lesser degree than do other groups (Delk & Ryan, 1977) or to the fact that the most common nontherapist subject population consists of college students. The latter explanation is supported by Tribich's (1977) finding of minimal sex role bias in the mental health judgments of either therapists or adult white-collar workers. It is also possible that the source of the differences lies in the experimental procedures and that the use of behavioral rather than trait descriptions and the use of multivariate dependent measures could affect the findings among clinicians.

Treatment Goals

Five studies have investigated the hypothesis that treatment goals tend to be sex role related (Table 3). These studies have generally failed to support this contention, although some specific goals may be sex typed.

Analogue Studies

Billingsley (1977) and B. J. Goldberg (1976) compared the treatment goals set for clients who varied in sex and problem. Neither study found stereotyped goals. Rather, the masculinity or femininity of goals depended on the type of problem: more feminine goals were recommended for potentially violent clients regardless of sex, and more masculine goals were recommended for withdrawn clients (Billingsley, 1977). Such goals suggest that the ideal person may be seen as a mixture of the masculine and feminine and that both extremes are undesirable.

Table 3
Studies That Investigated Therapy Goals

Study	Findings
Analogue studies	
Pringle (1973)	−
B. J. Goldberg (1976)	−
Billingsley (1977)	−
Field studies	
Fabrikant (1974)	−
Levy & Doyle (1974)	+

Note. + = supports hypothesis; − = does not support hypothesis.

Investigating counselor reactions to high and low dependence and achievement in male and female clients, Pringle (1973) found that both male and female counselors expressed a greater desire to change the behavior of low-achieving male clients than that of low-achieving female clients. In addition, male counselors recommended less change for dependent women and all high achievers than did female counselors.

Field Studies

Fabrikant (1974) surveyed patients in psychotherapy about their perceptions of their therapists' sex role goals for them. Only perceptions of male therapists were reported, and these perceptions generally failed to support the hypothesis of sex-role-related treatment goals. A sizeable minority (27%) of female patients, however, endorsed the statement that "male therapists encourage female patients to follow the role of wife/mother" (p. 101). On the other hand, they were not asked if their therapists discouraged such traits as independence and assertiveness.

Reporting on a drug rehabilitation program, Levy and Doyle (1974) state that "a stable relationship with a member of the opposite sex is important for a woman to complete the program while more credence is given to a male's realistic job plans" (p. 430). They neglect, however, to present any data relevant to the degree of importance of these two factors in making rehabilitation judgments, so no firm conclusions can be drawn.

Summary

Relatively little research has been conducted investigating treatment goals as a function of sex role stereotypes, and that tends to be generally negative. There are indications, however, that reactions to some specific behaviors (Pringle, 1973) and that some specific goals (Fabrikant, 1974) may be sex typed. This would still appear to be a fertile area for investigation.

Although adherence to a stereotype is generally not used as a treatment goal, it is possible that therapist-encouraged sex roles may have beneficial effects on some clients. Delk and Ryan (1977) note, for example, that Type A therapists, who stereotype the most, are also the most successful with schizophrenics. They suggest that these therapists provide structure and a "definitive role model of culturally valued and socially acceptable behavior" (p. 258) for a patient group characterized by ambivalence, confusion, and need for structure.

Conclusions

Nonprofessionals

Sex role stereotypes appear to be strong mental health cues for nonprofessionals. They hold stereotypic standards of mental health, and cross-sex-role traits and behaviors are seen as indicators of poor mental health and adjustment. When nonprofessionals evaluate reactions to stressful situations, however, it would appear that the situational context is a more important mental health cue than the reaction's sex role appropriateness.

Mental Health Professionals

Although the evidence concerning differential mental health standards for men and women indicates that clinicians share the sex role stereotypes of their lay contemporaries, there is little evidence that these stereotypes affect professional judgments or treatment goals. Although this apparent contradiction between studies of clinicians' attitudes and their behavior may be primarily due to the methodological limitations of the studies, as suggested by Davidson and Abramowitz (in press), there are two other possible explanations.

First, Gove (in press) has noted that a number of studies have found generally poorer mental health among women than among men, and he has concluded that if these data are accurate, then clinicians' differential mental health ratings of men and women could be more a reflection of reality than of stereotyped attitudes. In such a case one would not expect biased judgments or treatment goals, and the contradiction between the two sets of studies disappears.

Second, even given stereotyped attitudes, a contradiction does not necessarily arise. Social psychologists have long known that people's behavior relative to an object has little relationship to their attitudes toward the object. Fishbein and Ajzen (1975) suggest that a person's attitude toward a possible behavior directed at an object and the social norms that constrain that behavior determine the behavior more powerfully than does the person's attitude toward the object alone. Thus therapists' judgments of mental health and therapeutic goals may be more strongly determined by their attitude toward, and the norms surrounding, the imposition of traditional sex roles than by their sex role stereotypes alone. In the area of attitudes about specific behaviors, Fabrikant (1974) found that therapists valued the client's personal growth over traditional role enactment. This behavioral attitude is congruent with current therapeutic norms that stress the development and expression of the client's personal values, talents, and autonomy rather than passive social adjustment (cf. Corey, 1977; Orne, 1975). These norms have been emphasized especially strongly relative to woman clients (e.g., American Psychological Association, 1975; Rawlings & Carter, 1977). Under such conditions, sex role stereotypes may have little effect on therapist behavior.

The apparent exception to this rule concerns children. As previously noted, Feinblatt and Gold (1976) found that although therapist trainees were somewhat accepting of cross-sex-role behavior in children, they see a bleak future for them. This may reflect concern that the behaviors could lead to later homosexu-

ality, transsexuality, or social rejection (cf. Green, 1974; Rekers, Rosen, Lovaas, & Bentler, 1978).

Thus the influence of sex roles on psychotherapy appears to be more limited than some critics charge. This does not mean that the critics are necessarily wrong; it is more likely that therapists' attitudes have changed. Indeed, until recently, sex role congruence and mental health have been closely identified in the professional literature (e.g., Garai, 1970; Hurlock, 1974), and the change from this position may be attributable to the success of the women's movement in raising the consciousness of the psychotherapeutic establishment.

Reference Notes

1. Billingsley, D. *Sex-role stereotypes and clinical judgments: Negative bias in psychotherapy.* Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1976.
2. Frieze, I. H. *Changing self-images and sex-role stereotypes in college women.* Paper presented at the meeting of the American Psychological Association, New Orleans, La., September 1974.

References

Abramowitz, C. V., & Dokecki, P. R. The politics of clinical judgment: Early empirical returns. *Psychological Bulletin*, 1977, *84*, 460–476.

Abramowitz, S. I., Abramowitz, C. V., Jackson, C., & Gomes, B. The politics of clinical judgment: What nonliberal examiners infer about women who do not stifle themselves. *Journal of Consulting and Clinical Psychology*, 1973, *41*, 385–391.

Abramowitz, S. I., et al. Comparative counselor inferences toward women with medical school aspirations. *Journal of College Student Personnel*, 1975, *16*, 128–130.

American Psychological Association. Report of the Task Force on Sex Bias and Sex-Role Stereotyping in Psychotherapeutic Practice. *American Psychologist*, 1975, *30*, 1169–1175.

Anderson, M. Sex role stereotypes and clinical psychologists: An Australian study. *Australian Psychologist*, 1975, *10*, 325–331.

Aslin, A. L. Feminist and community mental health center psychotherapists' expectations for women. *Sex Roles*, 1977, *3*, 537–544.

Bem, S. L. The measurement of psychological androgeny. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 155–162.

Bem, S. L. On the utility of alternative procedures for assessing psychological androgeny. *Journal of Consulting and Clinical Psychology*, 1977, *45*, 196–205.

Berland, N. S. W. The effect of sex and sex-role on psychodiagnosis (Doctoral dissertation, Washington University, 1975). *Dissertation Abstracts International*, 1976, *36*, 4678-B. (University Microfilms No. 76-4742)

Bilick, J. G. The effect of patient gender on the clinical assessment process (Doctoral dissertation, University of Cincinnati, 1972). *Dissertation Abstracts International*, 1973, *33*, 3926-B. (University Microfilms No. 73-3823)

Billingsley, D. Sex bias in psychotherapy: An examination of the effects of client sex, client pathology, and therapist sex on treatment planning. *Journal of Consulting and Clinical Psychology*, 1977, *45*, 250–256.

Broverman, I. K., Broverman, D. M., Clarkson, F. E., Rosenkrantz, P. S., & Vogel, S. R. Sex-role stereotypes and clinical judgments of mental health. *Journal of Consulting and Clinical Psychology*, 1970, *34*, 1–7.

Burhenne, D. P. Female and male evaluations of sex-appropriate and sex-inappropriate sex-role stereotypes (Doctoral dissertation, Ohio State University, 1972). *Dissertation Abstracts International*, 1972, *33*, 910-B. (University Microfilms No. 72-20,947)

Chasen, B. Diagnostic sex-role bias and its relation to authoritarianism, sex-role attitude, and sex of the school psychologist. *Sex Roles*, 1975, *1*, 355–368.

Chasen, B., & Weinberg, S. L. Diagnostic sex-role bias: How can we measure it? *Journal of Personality Assessment*, 1975, *39*, 620–629.

Chesler, P. *Women and madness.* Garden City, N.Y.: Doubleday, 1972.

Coie, J. D., Pennington, B. F., & Buckley, H. H. Effects of situational stress and sex roles on the attribution of psychological disorder. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 559–568.

Collins, A. M., & Sedlacek, W. E. Counselor ratings of male and female clients. *Journal of the National Association of Women Deans and Counselors*, 1974, *37*, 128–132.

Corey, G. F. *Theory and practice of counseling and psychotherapy.* Monterey, Calif.: Brooks/Cole, 1977.

Costrich, N., Feinstein, J., Kidder, L., Marecek, J., & Pascale, L. When stereotypes hurt: Three studies of penalties for sex-role reversals. *Journal of Experimental Social Psychology*, 1975, *11*, 520–530.

Cowan, G. Therapist perceptions of clients' sex-role problems. *Psychology of Women Quarterly*, 1976, *1*, 115–124.

Davidson, C. V., & Abramowitz, S. I. Sex bias in clinical judgment: Later empirical returns. *Psychology of Women Quarterly*, in press.

De Beauvoir, S. *The second sex* (H. M. Parksley, trans.). New York: Bantam, 1967. (Originally published 1949.)

Delk, J. L., & Ryan, T. T. Sex role stereotyping and A–B therapist status: Who is more chauvinistic? *Journal of Consulting and Clinical Psychology*, 1975, *43*, 589.

Delk, J. L., & Ryan, T. T. A–B status and sex stereotyping among psychotherapists and patients. *Journal of Nervous and Mental Disease*, 1977, *164*, 253–262.

Derlega, V. J., & Chaikin, A. L. Norms affecting self-disclosure in men and women. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 376–380.

Fabrikant, B. The psychotherapist and the female patient: Perception, misperception and change. In V. Franks & V. Burille (Eds.), *Women in therapy*. New York: Brunner/Mazel, 1974.

Fabrikant, B., Landau, D., & Rollenhagen, J. Perceived female sex role attributes and psychotherapists' sex role expectations for female patients. *New Jersey Psychologist*, 1973, *23*(2), 13–16.

Feinblatt, J. A., & Gold, A. R. Sex roles and the psychiatric referral process. *Sex Roles*, 1976, *2*, 109–122.

Fischer, J., Dulaney, D. D., Fazio, R. T., Hudak, M. T., & Zinotofsky, E. Are social workers sexists? *Social Work*, 1976, *26*, 428–433.

Fishbein, M., & Ajzen, I. *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, Mass.: Addison-Wesley, 1975.

Garai, J. F. Sex differences in mental health. *Genetic Psychology Monographs*, 1970, *81*, 123–142.

Goldberg, B. J. Mental health practice as social control: Practitioners' choices of therapy goals as a function of sex of client, situations, and other practitioners' opinions (Doctoral dissertation, State University of New York at Stoney Brook, 1975). *Dissertation Abstracts International*, 1976, *36*, 5256-B. (University Microfilms No. 76-7570)

Goldberg, H. *The hazards of being male*. New York: Signet, 1976.

Gomes, B., & Abramowitz, S. I. Sex-related patient and therapist effects on clinical judgment. *Sex Roles*, 1976, *2*, 1–13.

Gove, W. R. The relationship between sex roles, marital status, and mental illness. *Social Forces*, 1972, *51*, 34–44.

Gove, W. R. Mental illness and psychiatric treatment among women. *Psychology of Women Quarterly*, in press.

Gove, W. R., & Tudor, J. F. Adult sex roles and mental illness. *American Journal of Sociology*, 1973, *78*, 812–835.

Green, R. *Sexual identity and conflict in children and adults*. Baltimore, Md.: Penguin Books, 1974.

Harris, L. H., & Lucas, M. E. Sex-role stereotyping. *Social Work*, 1976, *21*, 390–394.

Hill, C. E., Tanney, M. F., Leonard, M. M., & Reiss, J. A. Counselor reactions to female clients: Type of problem, age of client, and sex of counselor. *Journal of Counseling Psychology*, 1977, *24*, 60–65.

Hurlock, E. B. *Personality development*. New York: McGraw-Hill, 1974.

Hurvitz, N. Psychotherapy as a means of social control. *Journal of Consulting and Clinical Psychology*, 1973, *40*, 232–239.

Israel, A. C., Raskin, P. A., Libow, J. A., & Pravder, M. D. Gender and sex-role appropriateness: Bias in the judgment of disordered behavior. *Sex Roles*, 1978, *4*, 399–413.

Johnson, P. I. The relationship between sex-role stereotypes and concepts of mental health (Doctoral dissertation, Arizona State University, 1974). *Dissertation Abstracts International*, 1974, *34*, 5195-B. (University Microfilms No. 74-8969)

Komarovsky, M. *Dilemmas of masculinity: A study of college youth*. New York: Norton, 1976.

Kravetz, D. F. Sex role concepts of women. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 437–443.

Leifer, R. The medical model as ideology. *International Journal of Psychiatry*, 1970, *9*, 13–21.

Levy, S. J., & Doyle, K. M. Attitudes toward women in a drug abuse treatment program. *Journal of Drug Issues*, 1974, *4*, 428–434.

Lunneborg, P. W. Stereotypic aspects in masculinity-femininity measurement. *Journal of Consulting and Clinical Psychology*, 1970, *34*, 113–118.

Maslin, A., & Davis, J. L. Sex-role stereotyping as a factor in mental health standards among counselors in training. *Journal of Counseling Psychology*, 1975, *22*, 87–91.

Maxfield, R. B. Sex role stereotypes of psychotherapists (Doctoral dissertation, Adelphi University, 1976). *Dissertation Abstracts International*, 1976, *37*, 1914-B. (University Microfilms No. 76-22,816)

Nowacki, C. M., & Poe, C. A. The concept of mental health as related to sex of person perceived. *Journal of Consulting and Clinical Psychology*, 1973, *40*, 160.

Orne, M. T. Psychotherapy in contemporary America: Its development and context. In D. X. Freedman & J. E. Dyrud (Eds.), *American handbook of psychiatry* (Vol. 5, 2nd. ed.). New York: Basic Books, 1975.

Oskamp, S. *Attitudes and opinions*. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Pringle, M. B. The response of counselors to behaviors associated with independence and achievement in male and female clients (Doctoral dissertation, University of Michigan, 1973). *Dissertation Abstracts International*, 1973, *34*, 1627-A. (University Microfilms No. 73-24,659)

Rawlings, E. I., & Carter, D. K. (Eds.). *Psychotherapy for women: Treatment toward equality*. Springfield, Ill.: Charles C Thomas, 1977.

Rekers, G. A., Rosen, A. C., Lovaas, I. O., & Bentler, P. M. Sex-role stereotypy and professional intervention for childhood gender disturbances. *Professional Psychology*, 1978, *9*, 127–136.

Rosenkrantz, P., Vogel, S., Bee, H., Broverman, I., & Broverman, D. M. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 1968, *32*, 287–295.

Scott, W. A. Research definitions of mental health and mental illness. *Psychological Bulletin*, 1958, *55*, 29–45.

Shapiro, J. Socialization of sex roles in the counseling setting: Differential counselor behavioral and attitudinal responses to typical and atypical female sex roles. *Sex Roles*, 1977, *3*, 173–184.

Spence, J. T., Helmreich, R., & Stapp, J. Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 1975, *32*, 29–39.

Steinman, A., & Fox, D. J. Male-female perceptions of the female role in the United States. *Journal of Psychology*, 1966, *64*, 265–276.

Szasz, T. S. *The myth of mental illness*. New York: Harper, 1961.

Tennov, D. *Psychotherapy: The hazardous cure*. New York: Abelard-Schuman, 1975.

Terrill, M. J. *Sex-role stereotypes and conceptions of mental health of graduate students in counseling*. Urbana: University of Illinois, 1972. (ERIC Document Reproduction Service No. ED 101 255)

Thomas, A. H., & Stewart, N. R. Counselor response to female clients with deviate and conforming career goals. *Journal of Counseling Psychology*, 1971, *18*, 352–357.

Tribich, D. A. The influence of sex-role stereotypes and sex-role self-concept on judgments of mental health (Doctoral dissertation, Rutgers University, 1976). *Dissertation Abstracts International*, 1977, *37*, 5384-B. (University Microfilms No. 77-7290)

Zeldow, P. B. Effects of nonpathological sex role stereotypes on student evaluations of psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 304.

Zeldow, P. B. Sex differences in psychiatric evaluation and treatment: An empirical review. *Archives of General Psychiatry*, 1978, *35*, 89–93.

Received July 7, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1322–1338

Application of Biofeedback for the Regulation of Pain: A Critical Review

Dennis C. Turk
Yale University

Donald H. Meichenbaum
University of Waterloo, Ontario, Canada

William H. Berman
Yale University

The biofeedback literature for the regulation of pain is reviewed and found wanting on both conceptual and methodological grounds. In particular, studies on the use of biofeedback for the treatment of tension and migraine headaches and chronic pain indicate that biofeedback was not found to be superior to less expensive, less instrument-oriented treatments such as relaxation and coping skills training. The relative absence of needed control comparisons was noted, and the need for caution in promoting biofeedback was stressed. Suggestions for future research are offered.

During the past 10 years, there has been a proliferation of journal articles, books, research centers, workshops, and training programs devoted to biofeedback and the regulation of pain. The plethora of advertisements for biofeedback devices in medical and psychological magazines attests to the increasing popularity of these techniques as clinical tools. Perhaps the most telling evidence for the widespread popularity of this technique is the amount of coverage it has received in the media. In short, extraordinary public and professional attention has been given to the topic of biofeedback, often with spectacular claims made for the efficacy of these procedures for pain control. An examination of the research reveals a number of discrepancies and methodological problems that bring such glowing evaluations into question.

The purpose of the present article is to examine the empirical evidence that supports the relative efficacy of biofeedback techniques for the control of various pain syndromes. Because of the amount of data on this topic, the objective of this article is to provide a comprehensive rather than exhaustive coverage of the literature.

Requests for reprints should be sent to Dennis C. Turk, Department of Psychology, Box 11A Yale Station, Yale University, New Haven, Connecticut 06520.

What Is Biofeedback?

Biofeedback encompasses a wide variety of techniques that use biophysiological instrumentation to provide patients with information about changes in bodily functioning of which the person is usually unaware. The biofeedback training paradigm in general follows an operant learning framework. First, some body function that is to be brought under control is identified and assessed to provide a baseline of moment-to-moment changes. These changes are quantified and translated into information that is immediately fed back to the subject in the form of visual signals or auditory tones. The information regarding fluctuations in physiological activity provides subjects with data concerning the efficacy of their attempts to influence the specific physiological response. By trial and error the subject learns to recognize the subjective state and subtle internal changes associated with an alteration of physiological activity. As the subject becomes more proficient at controlling the signal, instrument sensitivity is adjusted to permit gradual shaping of the physiological response.

The physiological activity to be controlled varies as a function of the etiology of the symptoms to be regulated. Put simply, the intent of the biofeedback is to provide patients with information that will in turn enable

them to *voluntarily control* some aspect of their physiology that purportedly is causally linked to the pain experienced.

A number of important assumptions underlie the clinical use of biofeedback. One assumption is that the etiological variables and the pathophysiology of the pain to be controlled are known and can be subjected to voluntary control. A second assumption is that learned control of a bodily response is facilitated by information about activity in the relevant organ system. A third assumption is that through the biofeedback training, the patient will be able to recognize some of the situational factors that are related to the maladaptive physiological responding. A final assumption is that the skills learned during the biofeedback training will generalize to situations in the patient's natural environment, and as a result, these skills will be maintained over time and settings. That is, the patient will be able to engage in conscious control of relevant physiological responses outside the clinic or laboratory setting. Since most of the biofeedback training studies concerned with pain regulation have used patients with headaches (tension, migraine), this review first examines this literature and then examines other clinical pain populations.

Muscle Contraction Headache

The exact etiology of muscle contraction headache is unclear (Bakal, 1975). There is, however, a consensus that muscle contraction headaches (a) are an individual's response to psychological stress (American Medical Association, 1962; Dalessio, 1972; Martin, 1966) and (b) may result from excessive and sustained contraction of the frontalis (forehead), scalp, or neck muscles (Bakal, 1975; Martin, 1972). The traditional treatment of muscle contraction headaches usually entails symptomatic medication, for example, tranquilizers, muscles relaxants, or analgesics, and occasionally psychotherapy.

In 1954 Sainsbury and Gibson reported that the resting levels of frontalis electromyographic (EMG) activity were higher in patients with muscle contraction headaches than in normals. In 1969 Budzynski and Stoyva also demonstrated an association between frontalis EMG activity and tension in varying scalp and neck muscles. This association provided the basis for Budzynski, Stoyva, and Adler's (1970) suggestion that biofeedback would facilitate a patient's ability to attain "deep levels of relaxation" (p. 206) and would subsequently enable the client to "consciously control muscular tension" (p. 206). This led to a series of biofeedback studies (Budzynski et al., 1970; Budzynski, Stoyva, Adler, & Mullaney, 1973) in which tension headache patients were given frontalis EMG feedback.

In their initial report, Budzynski et al. (1970) provided five muscle contraction headache patients with frontalis EMG biofeedback training. Surface electrodes attached to the forehead provided information that was fed back by way of a tone, with the frequency of the tone proportional to the integrated EMG activity. Patients were instructed to attempt to keep the tone at the lowest frequency possible and were not provided with instructions as to how this might be accomplished. In addition, the patients were instructed to practice at home twice a day the skills that they had acquired during the feedback training. The amount of training received by the patients varied from 4 to 13 weeks, with two 30-minute sessions each week.

Budzynski et al. (1970) reported a steady decline in headache intensity and duration and EMG activity over the course of the training. A follow-up was conducted 3 months after the conclusion of the training. Muscle contraction headaches were eliminated in two patients and were reduced markedly in a third. For the remaining two patients, headaches returned shortly after the end of the feedback training.

Although the Budzynski et al. (1970) study demonstrates that patients can learn to reduce headache intensity as well as frontalis EMG activity, the effects cannot necessarily be concluded to be a direct result of the EMG feedback training. Because of the preliminary nature of the investigation, Budzynski et al. did not include any group to control for placebo or expectancy effects. Without such controls it is impossible to separate any active effect of the EMG training from nonspecific treatment effects associated with the

impressive biofeedback "ritual" (Miller, 1974; Shapiro & Surwit, 1976). Since the biofeedback treatment in Budzynski et al.'s study involved such components as EMG training, home practice, expectancy of alleviation of headaches, self-monitoring of stress-inducing situation, and a format that engendered a sense of control (cf. Epstein & Blanchard, 1977; Glass & Levy, in press), it is difficult to attribute improvement to the EMG feedback component. Another limitation of the Budzynski et al. study is the use of only a 3-month follow-up, since spontaneous fluctuations of headache incidence have been noted (Ostfeld, 1961). Given all of these shortcomings, the Budzynski et al. study provided impetus for further research.

In a better controlled study of biofeedback for the control of muscle contraction headache, Budzynski et al. (1973) compared the efficacy of frontalis EMG biofeedback, self-monitoring of headache intensity and duration with no feedback, and a noncontingent yoked feedback group. (The feedback signals were those produced by the true feedback group and were unrelated to any behavior on the part of the patients in this group.) Patients in the EMG and false feedback groups received 16 30-minute sessions of biofeedback training. The patients in both groups were instructed to practice at home twice a day the skills that they acquired during training.

At the end of training and at a 3-month follow-up, patients in the EMG feedback group had significantly reduced the intensity and duration of headache and frontalis muscle activity from the baseline period, whereas the self-monitoring and false feedback groups demonstrated no significant improvement. Interestingly, two patients in the biofeedback group who did not show significant reductions in headache intensity or duration both reported that they did not carry out the home practice, suggesting the potential importance of home practice in the biofeedback regimen.

Budzynski et al. (1973) contacted four of the six patients in the EMG feedback group 18 months after the completion of training. Three of the four patients reported that their headaches remained at very low levels with the fourth subject indicating that she had received some relief from headache activity following the training. This maintenance of improvement over such an extended period is impressive. Since the patients in the false-feedback and no-feedback groups were provided with true EMG feedback training after the 3-month follow-up, it is impossible to determine whether the continued improvement of the biofeedback group was a function of the training or of spontaneous fluctuations of headache incidence. Budzynski et al. did not provide any information as to whether the biofeedback subjects demonstrated reductions in EMG activity at the 18-month follow-up concordant with the reduction in headache incidence.

Although the second Budzynski et al. study (1973) was much better controlled, the design does not permit isolation of the contribution of the various components of the treatment regimen. The relevance of the EMG training to therapeutic gains is somewhat obscured by the instructions provided to the patients regarding the need to engage in home practice of skills acquired in the laboratory. The necessity of home practice has been addressed by several authors (e.g., Epstein, Hersen, & Hemphill, 1974; Haynes, Griffin, Mooney, & Parise, 1975; Wickramasekera, 1972), but its importance to biofeedback treatment is unresolved. Epstein et al. (1974) presented data from a single case study that supported the necessity of including home practice; on the other hand, Wickramasekera (1972) and Haynes et al. (1975) both reported substantial reductions in headache incidence resulting from biofeedback training that did not include suggestions for home practice.

All of the studies reviewed so far provide at least presumptive evidence that frontalis EMG biofeedback can reduce muscle contraction headache incidence. An important issue, however, is how useful biofeedback is in comparison to other less expensive approaches that are designed to regulate headache incidence. For example, several investigators (Epstein, Webster, & Abel, 1976; Fichtler & Zimmermann, 1973; Tasto & Hinkle, 1973) have reported that relaxation training without biofeedback training was sufficient to significantly reduce muscle contraction headaches. Another approach to alleviate muscle contraction headaches has

recently been used successfully by Holroyd, Andrasik, and Westbrook (1977). The therapeutic regimen used by Holroyd et al. was designed to train patients in a number of stress-coping skills to enable them to cope more adaptively with environmental and individual stylistic responses that instigated muscular tension. Further research should determine whether biofeedback techniques, which require sophisticated and expensive psychophysiological apparatus, are any more efficient or effective than more readily available and less expensive interventions.

Several recent studies (Chesney & Shelton, 1976; Cox, Freundlich, & Meyer, 1975; Haynes et al., 1975; Hutchings & Reinking, 1976; Holroyd et al., 1977) have addressed this issue by comparing the efficacy of frontalis EMG biofeedback with relaxation training or stress coping training. The data reported is equivocal but does not in general support the contention that biofeedback training is any more effective than various relaxation or cognitive control techniques.

In the Hutchings and Reinking (1976) study, three groups received EMG biofeedback training alone, relaxation training alone, or a combination of biofeedback plus relaxation training. All three groups were instructed to practice at home twice daily the skills that they had acquired. The groups who received the EMG training demonstrated significant reduction in headaches during a 28-day follow-up period compared to the relaxation training group. The EMG feedback and combined EMG feedback plus relaxation groups did not differ significantly in the incidence of headaches following training. Although patients in all three groups revealed reductions in frontalis muscle action potential following treatment, no significant between group differences were obtained, and no correlational data were reported between the amount of reduction in headache activity and the degree of EMG reduction. These data do suggest that biofeedback training is more effective than relaxation training; however, they also question the significance of frontalis EMG control for the reduction of muscle contraction headache.

Studies by Cox et al. (1975), Haynes et al. (1975), and Chesney and Shelton (1976) provide evidence that contradicts the Hutchings and Reinking (1976) findings. Although Cox et al. and Haynes et al. both report that EMG biofeedback was effective in alleviating headaches, groups who received the EMG training did not differ significantly from groups who received relaxation training. In the Chesney and Shelton (1976) study, only the relaxation training group demonstrated significant reductions in headache incidence, the EMG biofeedback group reduced headache incidence no more than a no-treatment control group.

In a recent study, Holroyd et al. (1977) used a therapeutic intervention that focused on altering maladaptive cognitive responses that were assumed to mediate the occurrence of muscle contraction headaches. Patients were provided with a rationale for treatment, which emphasized the function of specifiable maladaptive cognitions in the creation of subsequent disturbing emotional and behavioral responses (based on Beck, 1976 and Meichenbaum, 1977). Patients were encouraged to attribute their headaches to relatively specific cognitive self-statements rather than to situational or complex internal dispositions. A variety of stressful situations were identified, and patients were taught to focus (a) on the situational cues that trigger tension and anxiety for them, (b) on their response to these cues, (c) on their thoughts before becoming tense and after the development of tension, and (d) on the way in which these cognitions contributed to the tension headaches. Following this sequence, patients were instructed to deliberately interrupt the sequence of thoughts preceding their emotional response at the earliest possible point and to engage in cognitive control techniques incompatible with further stress and tension (e.g., cognitive reappraisal, attention deployment, fantasy).

This cognitive control regimen was employed with 10 tension headache patients who were compared to patients who received either biofeedback or no specific treatment. Training consisted of 8 biweekly sessions with a 15-week follow-up. At the termination of treatment and at follow-up, only the cognitive control group demonstrated substantial improvement on frequency, duration, and in-

tensity of headaches. Interestingly, only the biofeedback group demonstrated significant reductions in EMG activity. This latter finding raises the question of the assumed causal relationship between frontalis EMG activity and muscle contraction headaches.

Hutchings and Reinking (1976) also questioned the contribution of frontalis muscle tension in the development of muscle contraction headaches. The questionable relationship between EMG activity and experience of tension headaches is underscored further by Holroyd et al.'s (1977) data and the observation made by Haynes et al. (1975) that several of the patients in their study did not demonstrate a significant relationship between headache occurrence and EMG activity. Simply put, some individuals with extremely high frontalis EMG levels reported few or no muscle contraction headaches. Others with relatively low frontalis EMG reported frequent muscle contraction headaches.

Several recent studies have corroborated the independence between headache incidence and elevated EMG levels (Alexander, 1975; Cox et al., 1975; Coursey, 1975; Epstein & Abel, 1977; Holroyd et al., 1977). These recent findings are inconsistent with the significant relationship between headache incidence and EMG level ($r = .90$) reported by Budzynski et al. (1973). Cox et al. suggest that the discrepancies in the various studies may be accounted for by the sampling procedures of the EMG activity. Budzynski et al. (1973) recorded EMG levels during feedback sessions, whereas Cox et al. took EMG readings during pre- and posttreatment baseline sessions. But several other studies (Epstein & Abel, 1977; Haynes et al., 1975; Holroyd et al., 1977) that have sampled frontalis EMG levels in a method analogous to that of Budzynski et al. have not found a significant relationship between frontalis EMG and headache incidence, thus questioning Cox et al.'s explanation.

Other interpretations are possible. First, the etiology of muscle contraction headache may not result in high levels of frontalis EMG activity but rather in muscle contraction in other parts of the head, neck, and shoulders, with little generalization across the various muscle groups (Haynes et al., 1975). Second, changing frontalis muscle activity may not be sufficient for changes in self-report of headaches (Epstein et al., 1976; Holroyd et al., 1977). Third, the relationship between muscle contraction and muscle contraction headaches may not be isomorphic or unidimensional (Bakal, 1975). And finally, muscle tension changes may not always serve as the cue for reporting headaches (Epstein & Blanchard, 1977; Holroyd et al., 1977). In any case the evidence does not convincingly demonstrate that frontalis EMG reduction and self-report of tension headaches are concordant. Thus it seems premature to conclude that the positive effects of biofeedback approaches, when indeed they occur, are a function of increasingly voluntary control of frontalis muscle activity.

Table 1, which contains a summary of the studies that examined the relative efficacy of biofeedback in controlling muscle contraction headaches, places the biofeedback treatment procedure in some perspective. Perusal of Table 1 reveals that the majority of studies agree that frontalis EMG biofeedback is effective in the reduction of headache incidence, but when contrasted with alternative therapeutic interventions and no-treatment controls, these results are less impressive. In four out of seven studies, biofeedback training is no more effective than other approaches (e.g., relaxation, stress coping training) or no-treatment control groups. In addition, only four of nine studies reported any concordance between reductions in EMG levels and headache incidence.

Although the data are not conclusive, from a clinical perspective it appears that a number of psychologically oriented treatments (i.e., relaxation, biofeedback, cognitive control techniques) can be successfully used to ameliorate chronic muscle contraction headaches. It is thus unclear whether EMG biofeedback is more effective in relieving muscle tension than are other methods of relaxation, nor is it clear how patients use the skills acquired during training. Interviews with patients who have received biofeedback training reveal that they often generate cognitive instructions to themselves about what might relax them and then try out the approaches, using EMG feedback as a criterion (Hutchings & Reinking,

Table 1
Summary of Studies That Examined the Efficacy of Biofeedback in Regulating Muscle Contraction Headaches

Effect of frontalis EMG on headache activity			Comparison of effects of frontalis EMG activity and comparison groups[a]			Concordance between physiological activity and headache incidence		
Significant reduction	No significant reduction	Equivocal	EMG superior to comparison	EMG equal to comparison	EMG inferior to comparison	Concordant	Discordant	Equivocal
Budzynski, Stoyva, & Adler (1970)	Chesney & Shelton (1976)	Epstein & Abel (1977)	Wickramasekera (1972)[b]	Cox et al. (1975)	Chesney & Shelton (1976)	Budzynski et al. (1970)	Cox et al. (1975)	Hutchings & Reinking (1976)
Wickramasekera (1972)[b]	Holroyd, Andrasik, & Westbrook (1977)		Budzynski et al. (1973)	Haynes et al. (1975)	Holroyd et al. (1977)	Wickramsekera (1972)[b]	Haynes et al. (1975)	
Budzynski, Soyva, Adler, & Mullaney (1973)			Hutchings & Reinking (1976)			Budzynski et al. (1973)	Epstein & Abel (1977)	
Epstein, Hersen, & Hemphill (1974)						Epstein et al. (1974)	Holroyd et al. (1977)	
Cox, Freundlich, & Meyer (1975)								
Haynes, Griffin, Mooney, & Parise (1975)								
Hutchings & Reinking (1976)								

Note. EMG = electromyograph.
[a] All groups other than contingent frontalis EMG biofeedback training are treated as comparison groups. These include relaxation, noncontingent biofeedback, medication, and stress coping training (Holroyd et al., 1977).
[b] No statistical analyses were reported, thus we can only assume that the effects were significant.

1976). If an essential element of biofeedback training is the development of such cognitive and behavioral stratagems, then perhaps a more efficient and less expensive method of achieving relaxation would simply be to instruct patients in various methods of relaxation. The EMG biofeedback could then be used for a brief period to help the patient identify the most effective strategy. As Coursey (1975) asks, "what sort of relaxation technique is effective with what sort of people with what sort of problems in conjunction with what other procedures?" (p. 833).

It is clear that biofeedback training for the control of muscle contraction headaches directly addresses only the maladaptive physiological responses to stressful situations and ignores those psychological factors that initiate and contribute to such responses. As was noted previously, a three-stage process is generally believed to produce muscle contraction headaches: (a) The individual responds to psychological stress, which (b) may produce prolonged contraction of the muscles in the head, neck, or shoulders and which (c) may subsequently lead to the production of headache (American Medical Association, 1962; Bakal, 1975). To successfully treat muscle contraction headaches, a therapeutic regimen that addresses the first stage (response to psychological stress), such as the stress-coping training (Holroyd et al., 1977; Reeves, 1976), should be combined with treatments that focus on the second stage (maladaptive physiological responding). (See Turk & Genest, 1979, for an extensive review of such multifaceted treatment approaches with pain patients.) The failure of biofeedback procedures to specifically address individuals' maladaptive appraisals and behaviors in the natural environment may account for (a) the equivocal data, (b) for the relatively high number of patients who drop out of biofeedback treatment prematurely, as well as (c) the individual differences in the ability to benefit from feedback training (Lazarus, 1977; Meichenbaum, 1977; Turk & Genest, 1979).

In summary, the enthusiasm for the therapeutic application of frontalis EMG biofeedback for the control of muscle contraction headaches exceeds the available evidence of the efficacy of this approach as compared to other psychological approaches. Although biofeedback procedures have been shown to be useful, they should still be considered promissory at best, with many questions remaining unanswered.

Migraine (Vascular) Headache

The existing data on the pathophysiology of migraine headaches is sparse. The available physiological evidence suggests that migraine headache is associated with (a) excessive cranial vasculature responsivity and (b) autonomic nervous system instability (Bakal, 1975). The symptoms of migraine headaches are thought to be mediated through the autonomic nervous system and are often evidenced by increased blood flow in the head, which results in painful dilation and distention of the cranial arteries. A wide range of substances that produce vasoconstriction, including ergotamine tartrate, pituitrin, ephedrine, benzidrine, ephinephrine, and caffeine, have been shown to be more or less effective in relieving migraine, when administered prior to the establishment of edema (Dalessio, 1972).

Sargent, Green, and Walters (1972), who noted the association between migraine headache and cold extremities (Dalessio, 1972), hypothesized that the voluntary increase in finger temperature should be correlated with an increase in blood flow to the peripheral region and consequently a decrease of blood flow to the cranial region. In designing a treatment for migraine, these authors combined finger-temperature-warming biofeedback with autogenic training. Autogenic training (Schultz & Luthe, 1959) involves the simultaneous regulation of mental and somatic functioning by meditation on passive activities, a form of relaxation and self-instruction (e.g., "My mind is calm and quiet"; "My arms are heavy and warm").

The temperature biofeedback is designed to help the patient learn to vasodilate the skin blood vessels. Sensitive thermisters are attached to the forehead and index finger of the hand on the side most frequently affected by migraine. Information regarding the differential temperature between the forehead and index finger is fed back to the patient.

Patients are instructed to use the autogenic phrases to induce increased finger temperature. In essence the training is designed to teach the patient to abort the vasospastic phase of the migraine attack.

In one of the first attempts to use biofeedback techniques to control migraine, Sargent and his colleagues (1972) provided 75 patients who reported various types of headaches (the number of migraine patients was unspecified) with portable biofeedback devices and pages of autogenic phrases for daily home practice. After mastering hand warming, patients were instructed to practice temperature control at home on alternate days without the biofeedback apparatus.

On the basis of self-reports of the amount of analgesic medication used and frequency and intensity of headaches plus independent clinical ratings, Sargent et al. reported that 74% of the patients, as assessed by psychologists, benefited from the autogenic biofeedback training. However, of the total sample of patients, adequate clinical ratings were available on 62 (again, the number of migraine sufferers was unspecified), and pretraining data were available for only 32 migraine patients. Thus the authors can only confirm some degree of clinical improvement in 29% to 39% of the original sample of 75 patients. The Sargent et al. results can be contrasted to the results of a study conducted by Mitchell and Mitchell (1971) that reported significant improvements in migraine headache activity in 71% of patients treated with relaxation and other behavioral approaches.

The methodology of the Sargent et al. (1972) study has been seriously challenged by Blanchard and Young (1974). They note the following:

> The procedures for evaluating hand-warming are unsatisfactory for three reasons: (1) little or no data on results are given and it is reported that the posttreatment results do not reach statistical significance; (2) the treatment package itself is a mixture of several factors, suggestion, relaxation training and biofeedback training, any or all of which may have accounted for the results; and (3) no-treatment or attention-placebo treatment control groups were not included. (p. 586)

Several other studies have investigated the relative efficacy of biofeedback training for the amelioration of migraine headaches (Friar & Beatty, 1976; Kewman, 1978; Medina, Diamond, & Franklin, 1976; Mitch, McGrady, & Iannone, 1976; Turin & Johnson, 1976).

Mitch et al. (1976) employed Sargent et al.'s (1972) autogenic feedback training with 20 migraine patients. Training was conducted over a 12-week period. During the first month of training, patients were instructed to practice with the biofeedback device 30 minutes each day. In the second month subjects were instructed to practice daily and whenever they identified the onset of headaches. Finally, during the last month of training, patients were to practice 1–2 times per week, at the onset of headache, and on the evening of the day a headache had occurred. For 10 patients, follow-up was conducted 6 months after the completion of training.

During the training period, patients maintained records on four dependent measures: duration, frequency, and intensity of headaches and amount of medication taken. Patients were also asked to report on perceived changes of symptoms at the end of the treatment period. These ratings were contrasted with reports of headache and medication use during the 6 months prior to training. Judgment of the efficacy of this procedure was based on improvements on the subjective headache incidence and medication use reports. The authors reported that 65% of the patients improved on two or more of the dependent measures. At the 6-month follow-up, 9 out of 10 patients reported average to excellent improvement compared to that in the 6 months preceding treatment, which suggests an ability to control headaches over an extended period of time.

The criticisms offered by Blanchard and Young (1974) concerning the Sargent et al. (1972) study (failure to identify the effective components of the treatment, failure to employ adequate controls, and failure to provide adequate statistical analysis) can be applied to the Mitch et al. (1976) study as well. The Mitch et al. study has a number of additional flaws that make interpretation of the results tenuous. The efficacy of the autogenic training was based on retrospective self-reports of headache incidence and cannot be considered a valid baseline against which to compare treatment efficacy. No attempt

was made to measure forehead or finger temperature prior to, during, or following training. Thus determination of the influence of the autogenic feedback training on underlying physiological processes cannot be established. Selection of the 10 patients included in the follow-up is not specified by the authors, and thus sampling bias may be introduced. In sum, the study is inadequate on a number of grounds and does not permit any conclusion regarding the efficacy of autogenic feedback training per se.

Medina et al. (1976) reported retrospective data on 27 patients with migraine or mixed migraine and muscle contraction headaches. In this study, relaxation training, frontalis EMG training, hand warming, autogenic phrases, and home practice all contributed to the treatment. As in the two preceding studies (Mitch et al., 1976; Sargent et al., 1972), patients were provided with portable biofeedback devices for home practice. Booster biofeedback sessions were provided every 2 months. At follow-up (an average of 10.7 months following the completion of training) the authors indicated significant reductions in the number and severity of headaches and the use of medication in 13 of the patients. Nine of these patients with migraines (64%) and only 4 (30%) with mixed migraine and muscle contraction headaches benefited from the treatment. The same methodological weaknesses found in the previous studies apply to the Medina et al. study.

The relative contribution of the autogenic phrases, a component of the autogenic feedback training package that was used in these three studies (Medina et al., 1976; Mitch et al., 1976; Sargent et al., 1972), was examined by Turin and Johnson (1976). Seven patients suffering from migraine headache were trained in the finger warming procedures with the autogenic phrases omitted. Patients recorded the frequency and duration of headaches as well as the medication taken during a 4–6-week baseline and throughout the study. Following the baseline period, the biofeedback training was conducted over a 6–14-week period. During each session the first 25 minutes was devoted to acclimating the patients to the apparatus. Finger temperature was recorded during this habituation period and during the 20-minute biofeedback training period. In addition, patients were instructed to practice the temperature control skills twice a day and at the first sign of a headache. In contrast with the Sargent et al. and Mitch et al. studies, no portable device was provided for the patients to use in home practice.

Turin and Johnson (1976) provided three of their seven patients with temperature cooling training for 6 weeks prior to temperature warming training. The authors reasoned that if both cooling and warming produced significant effects, then a placebo expectancy hypothesis could explain the data.

All seven of the patients learned the peripheral warming task rapidly. Of the three patients who received the temperature cooling training, none showed clinical improvement under this condition. Subsequent to the temperature warming training, all patients reported significant reductions in both headache incidence and amount of medication taken. These data are consistent with case studies that used analogous temperature cooling training and temperature warming paradigm (Johnson & Turin, 1975; Wickramasekera, 1973).

The results of a recent study, however, indicate that the effects of finger-temperature cooling on headache incidence is equivocal. Kewman (1978) trained one group of migraine headache sufferers to raise their finger temperature and a second group to lower finger temperature; a third group received no training, but they self-monitored and recorded headache incidence. Surprisingly, all three groups showed significant reductions in headache incidence. Obviously, the reduction in headache incidence cannot be attributed unequivocally to the specific effects of temperature warming biofeedback training.

The Turin and Johnson (1976) study is a much better designed and controlled study than the Sargent et al. (1972), Medina et al. (1976), and Mitch et al. (1976) studies. Adequate baselines were obtained and attention-placebo procedures were used. The results of the Turin and Johnson study question the hypothesis that the treatment results were simply a function of the relaxation, home practice, or autogenic phrases. The Turin and Johnson study also questions whether home

practice on a portable feedback device is necessary. The Turin and Johnson study is limited, however, by the failure to include any follow-up data, by the absence of controls, and by the failure to provide correlations between temperature reduction and decrease in headache incidence.

Thus the exact relationship between peripheral finger temperature and migraine headache remains obscure. In the Turin and Johnson (1976) study, reduction in finger temperature accounted for at most 25% of the variance in headache improvement and was essentially unrelated to decrease in the number of headaches. Such findings suggest that nonspecific treatment effects associated with finger warming training may account for a significant portion of the outcome obtained with this treatment (Miller, 1974; Shapiro & Surwit, 1976).

Finally, in a carefully controlled study, Andreychuk and Skriver (1975) compared finger-temperature biofeedback, biofeedback for electroencephalograph (EEG) alpha enhancement, and self-hypnosis. The alpha enhancement group was included as an attention-placebo control that included the use of the biofeedback ritual (actually a rather powerful, potentially useful approach, used with some limited success by Gannon and Sternbach, 1971, in a single case study of migraine control). Andreychuk and Skriver obtained baseline headache information for each patient for the 6 weeks preceding training. Each of the three groups then received 10 45-minute therapeutic sessions, and patients were asked to practice the various skills at least twice a day between laboratory sessions. The pretreatment data were contrasted with the patients' headache incidence during the last 5 weeks of the 10-week training period. All three groups showed significant reductions in the incidence of migraine headaches, with no significant differences between the groups. Unfortunately no follow-up data was collected, and thus no determination of the maintenance of headache reduction can be made from the Andreychuk and Skriver study. The authors also failed to provide information about alterations in physiological activity as a function of any of the training regimens.

The hypothesis that the efficacy of the temperature feedback training is attributable to the specific biofeedback training does not receive support from the Andreychuk and Skriver (1975) study. All three treatment groups shared a number of components, namely, relaxation, home practice, an expectancy for relief, and a fostering of a sense of control, each of which may have engendered change. Interestingly, Andreychuk and Skriver assessed the hypnotic susceptibility of each patient and noted that the degree of headache improvement reported was strongly related to hypnotizability. This correlational data raises some intriguing possibilities regarding the effects of individual differences in suggestibility on the efficacy of biofeedback training. Further examinations of the relationship between such individual difference measures and biofeedback training need to be conducted before any firm conclusions can be drawn.

Friar and Beatty (1976) used a different approach to the control of migraine headaches with biofeedback techniques. In contrast to the other studies reviewed, Friar and Beatty did not use finger temperature biofeedback training as their experimental treatment. Rather, patients were trained to decrease pulse amplitude in either the head (experimental group) or a peripheral site (hand control group). The authors inferred that training to reduce pulse amplitude in the peripheral site should produce only nonspecific effects and would not influence migraine incidence.

Nineteen migraine sufferers were included in the Friar and Beatty (1976) study. Baseline ratings of the frequency and intensity of headache and amount of medication were obtained from all patients. The training consisted of eight sessions extended over a 3-week period. No specific instruction for regular home practice appears to have been offered. A ninth, no-feedback session was conducted to assess patients' ability to control response independent of feedback. In this session, patients were instructed to produce the vasoconstriction that they had learned in the laboratory whenever they became aware of developing headaches.

Table 2
Summary of Studies That Examined the Efficacy of Biofeedback in Regulating Migraine Headaches

Effect of biofeedback on headache			Effects of biofeedback compared to comparison groups/conditions[a]			Concordance between physiological activity and headache incidence		
Significant	Nonsignificant reduction	Equivocal	No comparison	Superior to comparison	Equal to comparison	Concordant	Discordant	Not reported
Wickramasekera (1973)[b]	Gannon & Sternback (1971)	Sargent, Green, & Walters (1972)	Gannon & Sternbach (1971)	Wickramasekera (1973)[b]	Andreychuk & Skriver (1975)	Johnson & Turin (1975)[b]	Kewman (1978)	Gannon & Sternbach (1971)
Andreychuk & Skriver (1975)		Sargent, Walters, & Green (1973)	Sargent et al. (1972)	Johnson & Turin (1975)[b]	Kewman (1878)	Wickramasekera (1973)[b]		Sargent et al. (1972)
Johnson & Turin (1975)[b]			Sargent et al. (1973)	Friar & Beatty (1976)				Sargent et al. (1973)
Friar & Beatty (1976)			Medina et al. (1976)	Turin & Johnson (1976)				Andreychuk & Skriver (1975)
Medina, Diamond & Franklin (1976)			Mitch et al. (1976)					Friar & Beatty (1976)
Mitch, McGrady, & Jannone (1976)								Medina et al. (1976)
Turin & Johnson (1976)								Mitch et al. (1976)
Kewman (1978)								Turin & Johnson (1976)

[a] All groups that were not designed to provide contingent biofeedback training, are treated as comparison groups. These include relaxation, noncontingent biofeedback, and medication.
[b] No signficant analyses were reported, thus we can only assume that the effects were significant.

Friar and Beatty (1976) reported no statistically significant differences between the two groups in the number of headache episodes or in the rating of mean intensity of headache episodes. The groups did differ significantly in the number of major migraine attacks (defined as lasting 3 hours or more); the experimental patients had fewer headaches. The authors suggest that the purpose of training in vasoconstriction is to "abbreviate the headache attack rather than prevent the onset" (p. 51).

No correlational data between pulse amplitude and headache incidence is presented by Friar and Beatty (1976), nor do they present any information about pulse amplitude before or after training, which limits the determination of the relationship between pulse amplitude and intensity of migraine. It is not clear if the migraines were reduced significantly during training or whether the reduction was maintained after treatment. No long-term follow-up was reported. Although the Friar and Beatty approach looks promising, it awaits replications with more attention to maintenance of effects.

Table 2 contains a summary of studies that used biofeedback to ameliorate migraine headaches. A pattern of results similar to that observed with muscle contraction headache is also evident in the biofeedback studies with migraine headaches. Although biofeedback techniques do seem to reduce the incidence of migraine headache, training in muscular relaxation produces similar results.

An underlying feature of these biofeedback studies is the hypothesis that learned vasomotor control is central to effective treatment of migraine headaches. Unfortunately the experimental uncertainties and the lack of quantifiable data, and necessary control procedures present serious difficulties, making endorsement of biofeedback methodologies tentative at best. One problem in particular stems from the failure to demonstrate that reduction in headache incidence is correlated with alteration of peripheral vasodilation or pulse amplitude, a premise on which the biofeedback training is based. The absence of such demonstrated relationships should qualify endorsements of the therapeutic efficacy of biofeedback treatments. Such caution is indicated by the absence of (a) a careful consideration of the placebo effect and expectancy demands and (b) comparisons with other less expensive and more readily available treatment approaches (e.g., relaxation and self-control procedures). Many important issues remain unresolved, for example, evaluation of the training parameters, duration of treatment effects, and the role of individual differences.

Chronic Pain Other Than Headache

In contrast to the host of studies that examined the effects of various biofeedback approaches for headaches, relatively few investigations have applied biofeedback techniques to other forms of chronic pain. Unsystematic case studies that reported the efficacy of biofeedback for chronic pain have been offered by Coger and Werbach (1975) and by Gentry and Bernal (1977). Other investigators have incorporated biofeedback into a variety of other procedures but have not assessed the contribution of the biofeedback training alone (e.g., Gottlieb et al., 1977; Newman, Seres, Yospe, & Garlington, 1978; Seres & Newman, 1976; Swanson, Swenson, Maruta, & McPhee, 1976; Khatami & Rush, Note 1).

Recently, two studies (Hendler, Derogatis, Avella, & Long, 1977; Melzack & Perry, 1975) have specifically examined the efficacy of biofeedback with groups of chronic pain patients. Hendler et al. used EMG biofeedback with chronic pain sufferers. They used frontalis EMG training for two reasons. First, Bonica (1974) had suggested that stress and anxiety could induce reflex muscle spasm, vasomotor changes, and local ischemia and thereby exacerbate pain syndromes that involved muscles, tendons, and reflex muscle spasms. Second, Budzynski and Stoyva (1969) had suggested that frontalis EMG relaxation was an indication of generalized muscle relaxation.

Thirteen patients suffering from a variety of pain syndromes were treated by Hendler et al. (1977) with five sessions of frontalis EMG feedback training. At a 1-month follow-up 6 patients reported that they were obtaining continued relief. The other 7 patients reported no benefit from the biofeedback training. No control procedures were used,

nor was specific information provided regarding the length of baseline and initial posttreatment and follow-up levels of EMG. Hendler et al. reported the absence of a significant correlation between muscle tension and reduction of pain. The absence of a relationship questions the conclusion that frontalis EMG training contributed to pain reduction. Hendler et al. concluded that

> the beneficial effects of biofeedback for these responders may be explained in terms of an increased sense of mastery over their environment, which resulted in a reduction of obsessive concern about their somatic problems and improvement in their self-esteem as a result of their increased environmental control. (p. 508)

Such a conclusion must be viewed only as a speculation, but a speculation that has been voiced by Lazarus (1977) and Meichenbaum (1977), who have also stressed the role of cognitive factors in biofeedback training.

In the second study that examined the efficacy of biofeedback in the reduction of chronic pain, Melzack and Perry (1975) compared the relative effects of EEG alpha biofeedback, hypnotic training, and a combination of both alpha and hypnotic training. The hypnotic training focused on increasing relaxation, energy levels, and mental calmness and on a reduction in the level of worry prior to the patients becoming upset. Six patients received the alpha feedback, 6 received self-hypnosis training alone, and 12 patients received the combined training. All three groups showed increased levels of alpha activity. The groups who received combined biofeedback and hypnosis training or hypnotic training alone demonstrated substantial reductions in pain compared to baseline. Patients who received the alpha training alone showed virtually no change in pain. These data suggest that the cognitive approach of hypnosis may be effective in reducing pain from unbearable to bearable levels but provide no support for the efficacy of EEG alpha biofeedback training as a tool in reducing pain. As Melzack (1975) suggested in the subtitle of his article that reviewed biofeedback approaches to reduce chronic pain, "Don't Hold the Party Yet." We concur!

Summary and Conclusions

What conclusions can be drawn from the studies that examined the efficacy of biofeedback for the regulation of pain?

1. The relationship between the experience of pain and the various physiological responses that biofeedback techniques are designed to control has not been established. Without the determination of some relationship between a physiological response and the experience of pain, the rationale for selecting a physiological function for voluntary control remains unclear.

2. Although it seems likely that most individuals can acquire some degree of voluntary control over autonomic functioning with biofeedback training, there are large individual differences among subjects. Investigations of the utility of biofeedback for pain regulation suffer from both "patient and treatment uniformity myths" (Kiesler, 1966, p. 129). Simply stated, investigators have treated all patients, regardless of individual differences, with ostensibly the same biofeedback therapy.

3. Many health care providers view biofeedback methodology as some unified therapeutic treatment. In actuality, biofeedback is a generic term for a wide range of approaches, which include some form of biological feedback that is intended to increase voluntary control of physiological responses. The question "Is biofeedback effective for regulating pain?" should be replaced by the questions "What combination of cognitive, behavioral, and biofeedback approaches would benefit which patients, with what symptoms, under what circumstances, and at what expense?" The present review questions the relative efficacy of biofeedback training in comparison with other more readily available methods. The promissory fanfare (and hoopla) that has accompanied biofeedback training has not yet been fulfilled. Some beginnings and some potential have been identified, but endorsements should be proffered with caution and qualification.

4. The active ingredients of biofeedback therapy have not been identified. The necessary and sufficient components of such a multifaceted approach as biofeedback have not been established.

5. Effective control procedures have not been used in the majority of studies. The inclusion of appropriate control groups is essential in the study of such conditions as migraine and muscle contraction headaches for several important reasons. Two sources of rival hypotheses that can be offered to explain the efficacy of biofeedback training are the regression-to-the-mean problem and the placebo effect. Miller (1974) argued that patients are much more likely to seek treatment when they are feeling worse. Since physiological systems tend to fluctuate between periods of exacerbation and those of amelioration, it is possible to obtain a sample of volunteers whose pain will show reduction due to spontaneous fluctuation, regressing toward the mean level. This point also underscores the necessity of incorporating extended follow-up periods to assess relative efficacy of biofeedback training.

Placebo effects are potent factors in any treatment of pain (Beecher, 1959; Evans, 1974; Ostfeld, 1961; Shapiro, 1963) and are especially potent in an impressive treatment such as biofeedback, with its complex mechanical equipment. The important potential role of placebo effects indicates the need for conducting double-blind studies that include credibility checks of patient expectancies (see Kazdin & Wilcoxon, 1976). The studies that have used feedback that involved temperature warming and cooling groups comprise a needed step in this direction.

6. There is relatively little information available concerning generalization of learning from the laboratory or clinic to the natural environment over extended periods of time. In many of the biofeedback studies, patients are instructed to become aware of the onset of headaches or some other physiological condition, but little or no attention is directed at what this entails. A more careful analysis of these processes should enhance the generalization process. A related issue is the need for more careful study of the dependent measures that have been employed in biofeedback studies. Three classes of measures have been used, namely, measures of physiological changes such as EMG, self-report ratings (e.g., headache intensity), and amount of medication. Much more concern should be directed at specifying the variables such as demand characteristics, scoring formats, and so forth, that influence the latter two measures. A much more careful analysis of these indices will provide implications for treatment interventions (see Frederiksen, Lynd, & Ross, 1978).

7. It is clear that pain conditions cannot be viewed as independent phenomena represented simply by maladaptive physiological responding. The particular physiological data gathered on any given day is a response determined by situational and diurnal variations and complex physiology and personality characteristics, as well as the particular environmental context. All of these factors need careful research consideration, particularly in view of the stress-producing potential of various environments (Insel & Moos, 1974). Research is needed to evaluate the interactions between situational effects and the person variables that contribute to individual differences in response to stress.

8. The focus of biofeedback training has been on increasing one's awareness of maladaptive physiological response by means of feedback and on developing voluntary control by conscious effort. As with any therapeutic regimen, biofeedback requires compliance over time. The question at the simplest level is how to motivate the patient to spend the necessary time practicing the desired behavior, especially when the novelty wears off. The problems inherent in convincing patients to continue to use prescribed medical or training regimens has been discussed by Agras (Note 2), Blackwell (1972), and Marston (1970).

9. Biofeedback training places the greatest emphasis on maladaptive physiological functioning. But this is a too restrictive view of pain. In pain syndromes, consideration must also be given to the patient's coping patterns and life style (Genest & Turk, 1979). Teaching voluntary control of physiological functioning may not be sufficient, since patients not only must control their physiology but they must be capable of dealing effectively with their environment (Shapiro & Schwartz, 1972). It may prove more feasible to consider biofeedback as an adjunctive technique to be used with other physiological and psychological approaches rather than as the sole

treatment modality (cf. Genest & Turk, 1979; Gottlieb et al., 1977; Mitchell & White, 1977; Turk & Genest, 1979; Khatami & Rush, Note 1). Training in the use of self-control strategies should enhance the utility of biofeedback techniques and should foster generalization beyond the relatively non-stress-producing, quiescent laboratory setting. (See Goldfried & Trier, 1974; Meichenbaum, 1977; Meichenbaum & Turk, 1976; Turk, 1977, for examples of such self-control interventions.)

10. The reviewed studies reveal a consistent lack of concern for the subjects' appraisals of the biofeedback techniques. What does the patient think of the training, and how does the patient use the skills acquired? These questions are rarely addressed in the literature. (cf. Meichenbaum, 1976, for a discussion of this issue.) Examination of patients who fail to benefit from biofeedback training and analyses of those patients who prematurely drop out of training would provide valuable information. The presentation of group data often tends to obscure individual differences in the ability to benefit from the training. Examination of single subject data would help to answer some of the questions raised previously.

In sum, the biofeedback studies reviewed do not yield consistent results. The evidence for the efficacy of biofeedback per se in reducing pain is marginal at best, resting mainly on case studies and poorly controlled research.

In drawing this conclusion, it is important to recognize that the diagnosis of various types of pain disorders is less than reliable. The lack of such reliability tends to limit the efficacy of any treatment, including biofeedback. Although the present review focused on chronic pain, similar conclusions can be offered about the questionable value of biofeedback with hypertension (Surwit, Shapiro, & Good, 1978) and even in the control of heart rate (White, Holmes, & Bennett, 1977). In some cases the evidence indicates that cheaper, more readily available relaxation and coping skills interventions are as effective or more effective than biofeedback training. Moreover, one cannot conclude from the aggregate of biofeedback studies that physiological feedback training per se is an indispensible (or even necessary) part of the therapeutic regimen. The only conclusion that seems warranted at this time is that one or a combination of the components of biofeedback training is an effective method of pain regulation for some patients in certain situations. Studies to date have dealt with diverse populations, a variety of biofeedback approaches that contain a number of potentially active components, and a wide array of research designs. Consequently, generalizations must be tentative. Current enthusiasm for biofeedback methodology must be tempered with an appreciation that the approaches are still experimental, with many issues as yet unexamined. Evidence for the widely acclaimed benefits of biofeedback is lacking, and biofeedback should be considered only a research tool at this time. Caution must be maintained to prevent the misapplication of biofeedback techniques.

Reference Notes

1. Khatami, M., & Rush, A. J. *A pilot study of the treatment of out-patirnts with chronic pain: Symptom control, stimulus control and social system intervention.* Paper presented at the meeting of the Association for the Advancement of Behavior Therapy, New York, December 1976.
2. Agras, W. S. *Application of behavior analysis to problems encountered in medical practice.* Paper presented at the meeting of the Association for the Advancement of Behavior Therapy, New York, December 1976.

References

Alexander, A. B. An experimental test of assumptions related to the use of electromyogram biofeedback as a general relaxation technique. *Psychophysiology*, 1975, *12*, 656–662.

American Medical Association. Report of the Ad Hoc Committee on the Classification of Headache. *Journal of the American Medical Association*, 1962, *179*, 717–718.

Andreychuk, T., & Skriver, C. Hypnosis and biofeedback in the treatment of migraine headache. *International Journal of Clinical and Experimental Hypnosis*, 1975, *23*, 172–183.

Bakal, D. H. Headache: A biophysical perspective. *Psychological Bulletin*, 1975, *82*, 369–382.

Beck, A. T. *Cognitive therapy and the emotional disorders*. New York: International Universities Press, 1976.

Beecher, H. K. *Measurement of subjective responses: Quantitative effects of drugs.* New York: Oxford University Press, 1959.

Blanchard, E. B., & Young, L. D. Clinical applications of biofeedback training. *Archives of General Psychiatry*, 1974, *30*, 573–589.

Blackwell, B. The drug defaulter. *Clinical Pharmacology and Therapeutics*, 1972, *13*, 841–848.

Bonica, J. J. Preface. In J. J. Bonica (Ed.), *Advances in neurology* (Vol. 4). New York: Raven Press, 1974.

Budzynski, T. H., & Stoyva, J. M. An instrument for producing deep relaxation by means of analog information feedback. *Journal of Applied Behavior Analysis*, 1969, *2*, 231–237.

Budzynski, T. H., Stoyva, J. M., & Adler, C. S. Feedback-induced muscle relaxation: Application to tension headache. *Journal of Behavior Therapy and Experimental Psychiatry*, 1970, *1*, 205–211.

Budzynski, T. H., Stoyva, J. M., Adler, C. S., & Mullaney, D. J. EMG biofeedback and tension headache: A controlled outcome study. *Psychosomatic Medicine*, 1973, *35*, 484–496.

Chesney, M., & Shelton, J. L. A comparison of muscle relaxation and electromyogram biofeedback treatments for muscle construction headache. *Journal of Behavior Therapy and Experimental Psychiatry*, 1976, *7*, 221–225.

Coger, R., & Werbach, M. Attention, anxiety, and the effects of learned enhancement of EEG in chronic pain: A pilot study in biofeedback. In B. L. Crue, Jr. (Ed.), *Pain: Research and treatment.* New York: Academic Press, 1975.

Coursey, R. D., Electromyogram feedback as a relaxation technique. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 825–834.

Cox, D. J., Freundlich, A., & Meyer, R. G. Differential effectiveness of electromyogram biofeedback treatments for muscle contraction headache. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 892–899.

Dalessio, D. J. *Woff's headache and other head pain.* New York: Oxford University Press, 1972.

Epstein, L. H., & Abel, G. G. An analysis of biofeedback training effects for tension headache patients. *Behavior Therapy*, 1977, *8*, 37–47.

Epstein, L. H., & Blanchard, E. B. Biofeedback, self-control, and self-management. *Biofeedback and Self-Regulation*, 1977, *2*, 201–211.

Epstein, L. H., Hersen, M., & Hemphill, D. P. Contingent music and anti-tension exercises in the treatment of a chronic tension headache patient. *Journal of Behavior Therapy and Experimental Psychiatry*, 1974, *5*, 59–63.

Epstein, L. H., Webster, J. S., & Abel, G. G. Self-managed relation in the treatment of tension headaches. In J. D. Krumboltz & C. E. Thoresen (Eds.), *Counseling methods.* New York: Holt, Rinehart & Winston, 1976.

Evans, F. J. The placebo response in pain reduction. In J. J. Bonica (Ed.), *Advances in neurology* (Vol. 4). New York: Raven Press, 1974.

Fichtler, H., & Zimmermann, R. R. Changes in reported pain from tension headaches. *Perceptual and Motor Skills*, 1973, *36*, 712.

Frederiksen, L., Lynd, R., & Ross, J. Methodology in the measurement of pain. *Behavior Therapy*, 1978, *9*, 486–488.

Friar, L. R., & Beatty, J. Migraine: Management by trained control of vasoconstriction. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 46–53.

Gannon, L., & Sternbach, R. A. Alpha enhancement as a treatment for pain: A case study. *Journal of Behavior Therapy and Experimental Psychiatry*, 1971, *2*, 209–213.

Genest, M., & Turk, D. C. A proposed model for group therapy with pain patients. In D. Upper & S. M. Ross (Eds.), *Behavioral group therapy.* Champaign, Ill.: Research Press, 1979.

Gentry, W. D., & Bernal, G. A. Chronic pain. In R. B. Williams & W. D. Gentry (Eds.), *Behavioral approaches to medical treatment.* Cambridge, Mass.: Ballinger, 1977.

Glass, C. R., & Levy, L. H. Perceived psychophysiological control: The effects of power versus powerlessness. *Journal of Personality and Social Psychology*, in press.

Goldfried, M. R., & Trier, C. S. Effectiveness of relaxation as an active coping skill. *Journal of Abnormal Psychology*, 1974, *83*, 348–355.

Gottlieb, H., et al. Comprehensive rehabilitation of patients having chronic low back pain. *Archives of Physical Medical Rehabilitation*, 1977, *58*, 101–118.

Haynes, S. N., Griffin, P., Mooney, D., & Parise, M. Electromyographic biofeedback and relaxation instructions in the treatment of muscle contraction headaches. *Behavior Therapy*, 1975, *6*, 672–678.

Hendler, N., Derogatis, L., Avella, J., & Long, D. EMG biofeedback in patient with chronic pain. *Diseases of the Nervous System*, 1977, *38*, 505–514.

Holroyd, K. A., Andrasik, F., & Westbrook, T. Cognitive control of tension headache. *Cognitive Therapy and Research*, 1977, *1*, 121–133.

Hutchings, D. F., & Reinking, R. H. Tension headaches: What form of therapy is most effective? *Biofeedback and Self-Regulation*, 1976, *1*, 183–190.

Insel, P. M., & Moos, R. H. (Eds.). *Health and the social environment.* Lexington, Mass.: Health, 1974.

Johnson, W. G., & Turin, A. Biofeedback treatment of migraine headache: A systematic case study. *Behavior Therapy*, 1975, *6*, 392–397.

Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, *83*, 729–758.

Kewman, D. C. *Voluntary control of digital skin temperature for treatment of migraine headaches* (Doctoral dissertation, University of Texas at Austin, 1977). *Dissertation Abstracts International*, 1978, *38*, 3400B. (University Microfilms No. 77-29, 053)

Kiesler, D. Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin*, 1966, *65*, 110–136.

Lazarus, R. S. A cognitive analysis of biofeedback control. In G. E. Schwartz & J. Beatty (Eds.), *Biofeedback: Theory and research.* New York: Academic Press, 1977.

Marston, M. Compliance with medical regimens: A review of the literature. *Nursing Research*, 1970, *19*, 312–323.
Martin, M. J. Tension headache, a psychiatric study. *Headache*, 1966, *6*, 47–54.
Martin, M. J. Muscle-contraction headache. *Psychosomatics*, 1972, *13*, 16–19.
Medina, J. L., Diamond, S., & Franklin, M. A. Biofeedback therapy for migraine. *Headache*, 1976, *16*, 115–119.
Meichenbaum, D. H. Toward a cognitive theory of self-control. In G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol. 1). New York: Plenum Press, 1976.
Meichenbaum, D. H. *Cognitive-behavior modification: An integrative approach*. New York: Plenum Press, 1977.
Meichenbaum, D. H., & Turk, D. C. The cognitive-behavioral management of anxiety, anger, and pain. In P. Davidson (Ed), *The behavioral management of anxiety, depression and pain*. New York: Brunner/Mazel, 1976.
Melzack, R. The promise of biofeedback: Don't hold the party yet. *Psychology Today*, 1975, *9*, 18–22.
Melzack, R., & Perry, C. Self-regulation of pain: The use of alpha-feedback and hypnotic training for the control of chronic pain. *Experimental Neurology*, 1975, *46*, 452–464.
Miller, N. E. Introduction: Current issues and key problems. In D. Shapiro (Ed.), *Biofeedback and self-control, 1973*. Chicago: Aldine Press, 1974.
Mitch, P. S., McGrady, A., & Iannone, A. Autogenic feedback training in migraine: A treatment report. *Headache*, 1976, *15*, 267–270.
Mitchell, K. R., & Mitchell, D. M. Migraine: An exploratory treatment application of programmed behavior therapy techniques. *Journal of Psychosomatic Research*, 1971, *15*, 137–157.
Mitchell, K. R., & White, R. G. Behavioral self-management: An application to the problem of migraine headaches. *Behavior Therapy*, 1977, *8*, 213–222.
Newman, R. I., Seres, J. L., Yospe, L. P., & Garlington, B. Multidisciplinary treatment of chronic pain: Long-term follow-up of low-back pain patients. *Pain*, 1978, *4*, 283–292.
Ostfeld, A. The study of migraine pharmacotherapy. *American Journal of Scientific Medicine*, 1961, *241*, 192.
Reeves, J. L. EMG biofeedback reduction of tension headaches: A cognitive skills-training approach. *Biofeedback and Self-Regulation*, 1976, *1*, 217–227.
Sainsbury, P., & Gibson, J. F. Symptoms of anxiety and tension and accompanying physiological changes in the muscular system. *Journal of Neurology, Neurosurgery and Psychiatry*, 1954, *17*, 216–224.
Sargent, J. D., Green, E. E., & Walters, E. D. The use of autogenic feedback training in a pilot study of migraine and tension headaches. *Headache*, 1972, *12*, 120–125.
Sargent, J. D., Walters, E. D., & Green, E. E. Psychosomatic self-regulation of migraine and tension headaches. *Seminars in Psychiatry*, 1973, *5*, 415–428.
Schultz, J. H., & Luthe, W. *Autogenic training: A psychophysiological approach in psychotherapy*. New York: Grune & Strattion, 1959.
Seres, J. L., & Newman, R. I. Results of treatment of chronic low-back pain at the Portland Pain Center. *Journal of Neurosurgery*, 1976, *45*, 32–36.
Shapiro, A. K. Psychological aspects of medication. In H. I. Lief, V. F. Lief, & N. R. Lief (Eds.), *The psychological basis of medical practice*. New York: Harper & Row, 1963.
Shapiro, D., & Schwartz, G. E. Biofeedback and visceral learning: Clinical applications. *Seminars in Psychiatry*, 1972, *4*, 171–184.
Shapiro, D., & Surwit, R. S. Learned control of physiological function and disease. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
Surwit, R., Shapiro, D., & Good, M. Comparison of cardiovascular biofeedback, neuromuscular biofeedback, and meditation in the treatment of borderline essential hypertension. *Journal of Consulting and Clinical Psychology*, 1978, *46*, 252–263.
Swanson, D. W., Swenson, W. M., Maruta, T., & McPhee, M. C. Program for managing chronic pain: 1. Program description and characteristics of patients. *Mayo Clinic Proceedings*, 1976, *51*, 401–408.
Tasto, D. L., & Hinkle, J. E. Muscle relaxation treatment for tension headaches. *Behaviour Research and Therapy*, 1973, *11*, 347–349.
Turin, A., & Johnson, W. G. Biofeedback therapy for migraine headaches. *Archives of General Psychiatry*, 1976, *33*, 577–579.
Turk, D. C. *A coping-skills training approach for the control of experimentally produced pain*. Unpublished doctoral dissertation, University of Waterloo, Ontario, Canada, 1977.
Turk, D. C., & Genest, M. Regulation of pain: The application of cognitive and behavioral techniques for prevention and remediation. In P. C. Kendall & S. D. Hollon (Eds.), *Cognitive-behavioral interventions: Theory, research, and procedures*. New York: Academic Press, 1979.
White, T., Holmes, D., & Bennett, D. Effect of instructions, biofeedback, and cognitive activities on heart rate control. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, *3*, 477–484.
Wickramasekera, I. E. Electromyograph feedback training and tension headache: Preliminary observations. *American Journal of Clinical Hypnosis*, 1972, *15*, 83–85.
Wickramasekera, I. E. Temperature feedback for the control of migraine. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, *4*, 343–345.

Received July 13, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1339–1349

Asymmetries in Processing Auditory Nonverbal Stimuli?

J. D. Craig
U. S. Army Human Engineering Laboratory, Aberdeen, Maryland, and
University of Delaware

Reports in recent psychoacoustical literature have suggested that pitch and rhythm properties of nonverbal stimuli may be processed differently by the two cerebral hemispheres. A review of this literature and pertinent background information generated the following conclusions: (a) Perception of pitch stimuli probably does not require differential cerebral processing. Only when some type of novel or complex time structure is generated in the stimulus presentation do the responses of subjects reflect a cerebral dominance effect. (b) The asymmetries demonstrated in the results of rhythm experiments closely parallel those found in verbal experiments. In addition, rhythm structure may also provide a framework for the synthesis and analysis of all incoming perceptual information. Further research on the properties of nonverbal auditory stimuli may add substantially to our knowledge of polysensory human information processing.

Recent psychoacoustical literature has suggested that pitch and rhythm properties of nonverbal stimuli may be processed differently by the two cerebral hemispheres. The present state of the art provides an unclear and sometimes conflicting picture of how either rhythm or frequency perception is related to cerebral dominance.

Some of the earliest studies in neurophysiology demonstrated that there were large differences in the functions of equivalent areas within the two cerebral hemispheres. In 1861 Paul Broca published a series of papers on language and the brain (cited in Geschwind, 1972). He identified the third frontal gyrus of the cerebral cortex as the area that when damaged, caused language deficiencies such as slow labored speech but did not affect comprehension of spoken language. Broca further reported that speech disorders occurred only when this area in the left hemisphere was damaged; damage to the corresponding structures in the right hemisphere did not produce corresponding clinical symptoms. The unequal specialization of the two halves of the brain became known as cerebral dominance.

Approximately 15 years later, Carl Wernicke described another language disorder (cited in Geschwind, 1972). Clinical symptoms included both quick articulate speech that was devoid of meaningful content and severe loss of understanding of spoken verbal material. Posthumous examination of the brains of these patients revealed lesions located between Heschl's gyrus and the angular gyrus in an area adjacent to the cortical auditory region. This area, now known as Wernicke's area, was located in the left hemisphere for most patients; damage to the equivalent area in the right hemisphere did not cause equivalent behavioral deficits.

Wernicke proposed a model of how this area interacted with Broca's area to provide normal speech capabilities. The key points of this model are as follows: (a) When a word is *heard*, it is conveyed to the auditory cortex, then relayed to Wernicke's area, where comprehension occurs. If the word is to be spoken, it is further conveyed to Broca's area via the arcuate fasciculus, a large band of fibers that connects the two regions. In Broca's area the spoken form of the word is aroused and passed on to the motor area that controls the muscles of speech. (b) When a word is *read*, output from

---

Requests for reprints should be sent to J. D. Craig, U. S. Army Human Engineering Laboratory, Behavioral Research Directorate, Aberdeen Proving Ground, Maryland 21005.

the visual cortex is relayed to the angular gyrus and further to Wernicke's area. In Wernicke's area the auditory form of the word is aroused and processing continues as described above. In terms of clinical value, this model has been effective in predicting which areas of the brain are involved in specific language disorders (Geschwind, 1972).

One of the most important findings in these early studies was that only one side of an individual's brain seems to be involved in language processing. For the vast majority of the patients studied, damage to the left hemisphere resulted in the language disorders described, whereas patients with equivalent damage to the right hemisphere did not develop language deficits. For a small percentage of the population, the opposite condition exists; that is, language abnormalities develop only when there is damage to the right hemisphere.

It was further noted that the great majority of right-handed persons were also left dominant for speech; that is, lesions in the left hemisphere of right-handed persons produced language disorders. However, cerebral dominance of the left-handed person was not nearly so well-defined; the left-handed person could be either right or left dominant for language functions. Currently accepted studies that correlate cerebral dominance with handedness indicate that 99% of all right-handed persons are left dominant for speech (Rossi & Rosadini, 1967), whereas for approximately half of all left-handers, language functions appear to be localized in the right hemisphere (Goodglass & Quadfasel, 1954). It is also accepted that lateralization of cerebral dominance is less clearly defined in left-handed persons. In many behavioral studies this correlation between handedness and cerebral dominance has been used as a convenient means for assuming cerebral dominance in a subject. This assumption is probably more valid for right- than for left-handed persons.

Since the landmark studies by Broca and Wernicke, a great deal of effort has been expended in attempting to define the exact nature of cerebral dominance. A short, selective review of both physiological and behavioral studies is provided here, with emphasis on the research that provides a perspective for current studies of asymmetrical processing of nonverbal auditory stimuli.

Physiological Studies

The central auditory system in both humans and animals is physically a bilaterally projecting system. However, most of the stimulation received by either ear is conveyed to the opposite temporal cortex via the strong contralateral auditory pathways. Much of the information that does reach the ipsilateral cortex is derived from input that has traveled along the contralateral pathway and been redirected back to the side of original stimulation. The secondary and efferent pathways are so complex that many details concerning them are still uncertain (Carpenter, 1976). Nevertheless, it is clear that the auditory pathways function as a primarily contralateral system.

Dominance is usually associated with the cerebral hemispheres only. Although much information on the anatomy of the brain has been derived from animal studies, research on cerebral asymmetries using animal subjects is still considered by some writers to be of questionable value because many of the behavioral manifestations of cerebral dominance, like language function, are absent in animals. Summarizing the evidence presented by several investigators concerning lateralization in animals, in particular "pawedness" in cats, rats, and primates, Jung (1962) concluded that "real hemispheric dominance does not occur in any of these infrahuman species" (p. 268).

Since that time, the most compelling evidence for cerebral dominance in animals is found in the experiments of Nottebohm (1970). He found that the pattern of song in chaffinches and canaries is largely destroyed if the left hypoglossal nerve is lesioned but remains undisturbed when the right hypoglossal nerve is cut. His inference was that birdsong is primarily controlled by the left side of the brain in these birds.

In addition to the uncertainty that exists concerning the value of animal experiments in this area, there has been some question until recently whether there are structural asymmetries in the human brain that can be correlated with the functional asymmetries noted earlier. For many years it was generally ac-

cepted that the language dominance found in humans is not associated with any significant differences in anatomy between the right and left hemispheres. Von Bonin (1962) reviewed a large number of studies on this subject and concluded that "these morphological differences are, after all, quite small. How to correlate these with the astonishing differences in function on the left side, is an entirely different question" (p. 6).

In 1968 Geschwind and Levitsky published the results of their postmortem examinations of 100 adult human brains. The hemispheres were divided and each hemisphere was sectioned along the plane of the sylvian fissure to expose the upper surface of the temporal lobe. They found that the planum temporale was larger on the left for 65% of their specimens and on the right for 11% ($p < .001$), but 24% of the specimens showed equality of the two sides. This is important because the planum temporale contains the auditory association cortex. (The primary auditory cortex is located in Heschl's gyrus.) In the dominant hemisphere, these regions of the auditory association cortex are the classical Wernicke's area. Unfortunately, no information was available about the cerebral behavioral dominance or handedness of the living subjects. However, since 93% of the adult population is right-handed, the authors concluded with some small tolerance that they were probably working with left-dominant specimens. Their conclusion was as follows:

> Our data show that this area is significantly larger on the left side, and the differences observed are easily of sufficient magnitude to be compatible with the known functional asymmetries. (Geschwind & Levitsky, 1968, p. 187)

Recently, Yeni-Komshian and Benson (1976) have reported that chimpanzee brains have a similar asymmetry but to a lesser degree than do human brains. The brains of rhesus monkeys did not show any significant differences between the right and left temporal lobes. Since there have been demonstrations of some language capability among chimpanzees, these investigators suggested that neuroanatomical asymmetry may be a prerequisite for language functions.

To date, however, the Geschwind and Levitsky (1968) research is the most definitive study available on the anatomical asymmetries that may be associated with the functional differences found in human clinical and behavioral studies.

Behavioral Studies

The behavioral studies of the role of cerebral dominance in auditory perception are almost exclusively dichotic-listening experiments. Dichotic listening is a technique in which different inputs are simultaneously delivered to the right and left ears.

Kimura (1961) adapted the dichotic method to the study of the cerebral dominance effect in audition. She showed that when two digits are simultaneously presented to the two ears of a normal subject, digits arriving at the ear contralateral to the dominant hemisphere are more readily recognized than those arriving at the ipsilateral ear. At the end of her study, she concluded that the crossed auditory pathway from the contralateral ear to the speech hemisphere is more effective than the slightly smaller uncrossed pathway from the ipsilateral ear and that the dominant temporal lobe is more important than the nondominant temporal lobe in the perception of speech.

In a later study Kimura (1967) noted that in auditory studies, a cerebral dominance effect is evident only when there is simultaneous input to the two ears, that is, when dichotic presentation is used. Laterality, or ear superiority, is not evident when identical material is presented monaurally. If there is only one source of stimulation at a time, each ear performs equally well. Summarizing the difference between the dichotic and monaural studies, Kimura proposed that the cerebral dominance effect is the result of competition between simultaneous inputs to opposite cerebral hemispheres. When there is dichotic stimulation to the two hemispheres, competition between the coincident stimuli automatically occurs. Superior responses to the stimuli presented to the dominant hemisphere result from the conflict between the disparate perceptions of the two hemispheres.

This conclusion went somewhat beyond the data presented, and shortly thereafter several efforts were made to disprove it. In particular, two major objections were tested. The first

objection to Kimura's (1967) hypothesis focused on the role of memory. Laterality in response might be due to asymmetries in recall rather than to asymmetries in perception. The objective, then, was to separate the perceptual from the storage or response phases of the dichotic listening method.

Bryden (1967) conducted a series of experiments to study this issue. He suggested that the material from the ear that was reported first would be identified more accurately than material from the other ear. This would occur because the time elapsed decreases the accuracy of memory for the second channel. Therefore, a tendency to consistently report material from a preferred ear would account for the laterality effect, even if initial perception of material to both ears were equal. He examined the difference between free recall, in which the subject was allowed to report material from either ear at will, and ordered recall, in which the subject was required to report all material from one or the other ear first.

He found a high correlation between right-ear advantage in free recall and right-ear dominance in ordered recall. His data indicated that material presented to the right ear was more accurately identified than material presented to the left ear. (He had primarily used right-handed subjects.) This was true for all rates and list lengths investigated. When the two ears were compared as channels of immediate recall, the right ear was superior to the left. Also, the right ear was better as a storage channel than the left. In addition, there was a general tendency to report the right ear first. He concluded that the data supported the notion that right-ear superiority is due to a perceptual difference rather than to an order effect. This conclusion clearly supports Kimura's (1967) hypothesis that the cerebral dominance effect is a result of competition between simultaneous inputs to the two cerebral hemispheres.

The second objection to Kimura's (1967) hypothesis concerned the role of attention in the results of dichotic listening studies. Bryden (1969) tested a hypothesis that the laterality effect obtained by Kimura was due to division of attention rather than to competition of simultaneously arriving stimuli. In the first experiment subjects listened to monaural stimuli but had no prior knowledge of which ear would receive the next stimulus (monaural presentation with division of attention). Responses in these conditions showed no laterality whatsoever.

Bryden (1969) further tested his subjects using dichotic input under two different conditions. In the first condition, subjects were told to which ear to attend; therefore, they ostensibly were attending to only one channel while receiving competitive stimulation through both ears. In the second condition, subjects were not told to which ear they should attend; this condition offered both stimulus competition and division of attention. In both of these conditions a statistically significant laterality effect was obtained. To summarize, regardless of instructions or deliberate direction of attention, lateralization of response occurred in the dichotic listening paradigm. Bryden concluded that these results supported Kimura's (1967) hypothesis that the laterality effects obtained in dichotic listening experiments are due to signal competition rather than to attention factors.

Another variation of the dichotic listening studies of cerebral asymmetry is the research in which pathological subjects have been used. Milner, Taylor, and Sperry (1968) found that right-handed commissurotomized patients (those who have surgical disconnection of the cerebral hemispheres because of epilepsy or other reasons) could not report verbal input to the left ear if a different verbal input was simultaneously delivered to the right ear. However, all known auditory pathways remained intact and the subjects could report with total accuracy monaural input to either ear.

The results of this study were duplicated by Sparks and Geschwind (1968). A review of the data led them to propose another model for dichotic auditory asymmetries, which incorporated Kimura's (1967) model but suggested in addition a callosal auditory pathway between the two cerebral hemispheres. This model could account for the cerebral dominance effect evident in normal subjects and the left-ear suppression by right-handed commissurotomized patients in dichotic studies. The main points of this model were as follows:

1. In dichotic listening contralateral ear input virtually suppresses ipsilateral input.

2. There is competition for report by the left-hemisphere speech system between information arriving directly from the right ear via the contralateral pathway and information from the left ear.

3. Since information from the left ear has also traveled along a contralateral pathway to the right hemisphere, it must in addition be projected to the left hemisphere for report. This projection probably involves a callosal pathway.

In a further study, Sparks, Goodglass, and Nickel (1970) used this model to explain data gathered from left-brain-injured aphasic patients and right-brain-injured nonaphasic patients. The right-brain-damaged group could not report the signals received by the left ear after listening to dichotic verbal stimuli. However, the left-brain-damaged group was divided between those who experienced inhibition of right-ear input and those who experienced inhibition of the left-ear input. One possible explanation for these results was that competition between signals received by both ears occurs exclusively in the left hemisphere. Therefore, they revised the earlier model to state that only damage to the left hemisphere can affect information from either the contralateral or ipsilateral ear.

No review of the literature that deals with cerebral asymmetries and audition is complete without some mention of the studies that have concentrated on the functions of the minor hemisphere. Originally this work began by drawing an analogy from studies of cerebral asymmetries in vision. These studies showed that damage to the nondominant temporal lobe produced impaired performance on many visual, nonverbal tasks. A natural extension of this work was to determine if such a division of function also existed in the auditory modality.

Milner (1962) examined the effects of temporal lobectomy on nonverbal auditory discriminations. Her subjects were left dominant for speech; in addition, each subject had a lesion in either the right or left temporal lobe. These subjects responded to the Seashore Measures of Musical Talents, which included tests for pitch, loudness, rhythm, time, timbre, and tonal memory. Her data showed that the group with right temporal lesions made more errors than the group with left temporal lesions. The difference between the two groups was strongest for tonal memory and timbre. This research definitely indicates that the right hemisphere is strongly involved in processing certain types of musical sounds in left-dominant subjects.

Later, Kimura (1964) verified these results using normal subjects. Her rationale was that if the nondominant hemisphere is more strongly involved in certain musical abilities than the dominant, then normal subjects who show right-ear advantage for verbal materials should also show a left-ear advantage for musical material. This reasoning was based on a knowledge of the strong contralateral auditory pathways and was consistent with her model for dichotic auditory processing.

In preliminary tests, subjects listened to different numbers of clicks presented simultaneously to both ears. They were required to report the number of clicks presented to each ear. The subjects responded with a small but nonsignificant bias in favor of the left ear. She then presented melodic patterns dichotically. Left-ear melodies were reported correctly significantly more often than right-ear melodies ($p < .01$). Kimura (1964) concluded on the basis of these data and Milner's (1962) studies that the difference in function between major and minor hemispheres is along a verbal–nonverbal dimension. Kimura also noted that this asymmetry is obtained only in dichotic listening conditions.

Kimura's model provided a simple, easily applied paradigm for explaining large groups of existing data and experimental results. One of the first aspects of the model to receive attention concerned the meaningfulness of stimuli. Curry (1967) investigated this problem in a three-condition task using dichotic words (meaningful verbal), dichotic nonsense syllables (nonmeaningful verbal), and dichotic environmental sounds (nonverbal). His subjects were instructed to identify both stimuli in a free-recall paradigm. They obtained higher scores for right-ear stimuli when both words and nonsense syllables were used but higher left-ear scores with the nonverbal stimuli. This study shows that meaningfulness is not critical for the functional division ob-

tained in this and in other studies and supports Kimura's model.

Shortly after this time, research began to appear which indicated that the situation is not a simple dichotomy of function. Studdert-Kennedy and Shankweiler (1970) presented data which suggest that consonants are processed by the left hemisphere, whereas vowel sounds are processed by both hemispheres. The stimuli were spoken consonant–vowel–consonant syllables presented in dichotic pairs. For any pair of stimuli, only the initial consonants, the final consonants, or the vowels differed. Subjects were tested separately for each of the three types of stimuli. They were told to report both initial consonants in the dichotic pair, both final consonants, or both vowels. Significant right-ear advantages were obtained for the initial ($p < .001$) and final ($p < .01$) consonants. Data showed mixed ear superiority for the vowel sounds, which suggests that both hemispheres are involved in speech analysis.

Studies that investigated the processing of nonverbal dichotic stimuli also report a complex division of function. These studies suggest that different acoustical attributes of nonverbal stimuli are differentially processed by the cerebral hemispheres. Spellacy (1970) found a significant left-ear advantage for dichotic melodies but found no significant difference between ears for timbre, temporal, or frequency patterns. Stimuli used for the melodies test were unfamiliar violin solo melodies. Frequency patterns were composed of four 500-msec consecutive tones. Each tone was of a different frequency and all tones were between 440 Hz and 880 Hz. Temporal stimuli were tone pulses arranged in Morse code patterns. Timbre stimuli consisted of single notes played on a pipe organ, using varying combinations of pipes. After listening to the dichotic test stimuli, subjects listened to binaural identification stimuli and then reported whether the identification stimulus matched either of the test stimuli.

Gordon (1970) also attempted to separate the different acoustical qualities found in many nonverbal stimuli. He devised two tests for his subjects. The first consisted of melodies largely devoid of timbre and chordal properties. The second test consisted of electric organ chords that were rich in timbre. The first test therefore varied pitch over time to produce melodies; the second test avoided melodic sequence. Gordon's results, in contrast with those of Spellacy (1970), showed no asymmetry in recognizing and reporting the melodies but showed a strong left-ear advantage in the chords test.

One of the things that makes this report by Gordon interesting is his explanation of these results. He noted that it is impossible to rule out a hypothesis that rhythm may be lateralized to the left hemisphere:

> [A] predominance of rhythmic function in the left hemisphere coupled with the function of pitch discrimination in the right hemisphere could produce the observed results and confound the conclusions. The subjects would recognize unique rhythms in some trials and unique melodic changes in others so that the overall performance [on the melodies test] would show no asymmetry. (p. 395)

Pitch Processing

Deutsch (1974) also reported some highly unusual data. She described perceptions of certain dichotic pitch patterns as "auditory illusions" because they differed considerably from the actual physical stimuli presented. She reported that perception of these illusions was related to the handedness and presumed cerebral dominance of the subjects. The illusion was based on opposing octave pitch patterns and is hereafter called the "octave illusion."

The octave illusion occurred under the following conditions. A sequence of tones was presented to one ear, alternating in frequency between 400 Hz and 800 Hz. Tone duration was 250 msec and there were no intervals between tones. The opposite ear received an equivalent sequence simultaneously at equal amplitude. However, the sequences to the two ears began on opposite frequencies, so that simultaneously, a 400-Hz tone was presented in the right ear and a 800-Hz tone was presented to the left ear.

None of the 86 subjects accurately reported the stimuli. Right-handed subjects had a tendency to hear a single tone oscillating from ear to ear; the pitch of the tone also oscillated from one octave to the other. Approximately half of the left-handed subjects reported this

pattern, too. However, 39% of the left-handed subjects reported complex perceptions, for example, two alternating pitches in one ear with a third pitch intermittently in the other or two alternating pitches oscillating from ear to ear and two further alternating pitches localized at the back of the head. Right- and left-handed subjects differed significantly in the relative distribution of their percepts.

Deutsch (1974) further reported that right-handed subjects had a significant tendency to report the high tones in the right ear and the low tones in the left ear. However, she indicated that this localization pattern often reversed with continued listening, much as perceptions of ambiguous visual patterns reverse.

Not all investigators agree on the role of cerebral asymmetries in the processing of pitch information. In particular, Efron and Yund (1976) reported data in which subjects favored one ear or the other in reporting certain dichotic chords but found no correlation between the favored ear and the cerebral dominance of their subjects. They described results of this type as an "ear dominance effect."

A typical example of the type of stimulation used by Efron and Yund (1976) in many experiments was as follows: One ear received a 1900-Hz tone for 320 msec; after a 4-msec interval, a second 320-msec pure tone of 1500 Hz was presented (high–low pattern). The opposite ear received simultaneous tones, with the order of the frequencies reversed (low–high pattern). Subjects reported whether the pitch of the first chord was higher or lower than that of the second. Describing their results across many experiments, Efron, Dennis, and Yund (1977) stated that

> although both frequencies comprising the dichotic chord are heard by all subjects with normal hearing, only one-third of the subjects hear the two frequencies with approximately equal salience. For another third of the population, the pitch mixture of the dichotic chord is unequivocally dominated by the frequency delivered to the left ear; for the final third of subjects, the frequency of the tone presented to the right ear dominated the pitch mixture of the dichotically presented chord. (p. 538)

The distribution of these data does not correspond to any known factor related to cerebral dominance or handedness. In addition, Efron and Yund (1976) investigated the relation of ear dominance to the handedness of their subjects. They found that there was no correlation with handedness when dichotic tones of about 1700 Hz were presented to the subjects.

They further suggested that the difference between their data and those reported by Deutsch (1974) might be due to the higher frequency of their stimuli. Deutsch used stimuli in the 400 to 800 Hz range, which is within the frequency range of speech vowel sounds. Although Stevens and House (1972) summarized data which show that important features of the speech frequency envelope range from 300 to 3000 Hz, Efron and Yund (1976) concluded that "correlation between ear dominance for pitch and handedness (hemispheric dominance for speech) exists only in the frequency band which carries speech information" (p. 898). We must assume that they were referring to the frequency band for vowel sounds.

Such a hypothesis did not go unnoticed. Christensen and Gregory (1977) reported an experiment designed to examine the ear dominance effect in the speech vowel frequencies. They used 400- and 800-Hz tones in an experimental paradigm that closely resembled that used by Efron and Yund (1976). Subjects received two consecutive pairs of tones. Tones lasted 250 msec; one channel consisted of a 400-Hz tone followed by an 800-Hz tone, whereas frequencies were reversed for the second channel. Most subjects perceived only a single tone from each of the dichotic pairs. This is similar to Deutsch's (1974) results. However, in agreement with Efron and Yund's results, they found no evidence of any type of a cerebral dominance effect in the responses.

From the previous review it is apparent that there is in the literature a dichotomy pertaining to pitch processing. When some dichotic pitch stimuli are presented, asymmetries in response show a relationship with handedness; when other dichotic stimuli are used, no relationship between handedness and responses is found. Since both sets of stimuli were designed to study pitch processing, it is unclear why this happens. It is possible that one of the paradigms may be unwittingly contaminated by the inclusion of some factor other than simple pitch discrimination. If this is true, then there

is a possibility that the presence of a cerebral dominance effect in one group of data and an unrelated ear dominance function in the other group is due to some factor other than differential pitch processing.

There are two major dimensions present in any acoustical pattern: pitch and time. The two dimensions are probably inseparable on an absolute basis. When we speak of frequency, we describe sound as a certain number of cycles per second. When we speak of any type of acoustical time or rhythm pattern, the elements of that pattern are composed of specific frequency profiles. In dichotic listening experiments, study of one dimension has been accomplished by exactly matching the other dimension in both ears. For example, in both the Deutsch (1974) paradigm and the Efron and Yund (1976) paradigm, the on–off times of stimuli to opposite ears are exactly synchronized, whereas only the frequency of the stimuli presented to each ear varies.

However, there is another time-related variable that is introduced in comparisons of the two sets of stimuli: the number of frequency changes or transitions in each pattern. Halperin, Nachson, and Carmon (1973) found a shift in ear advantage related to the number of frequency and duration transitions within temporally patterned nonverbal stimuli. They found that as the complexity of the pattern increased (the number of transitions within the pattern), there was a shift from left- to right-ear superiority with right-handed subjects.

In the Deutsch (1974) paradigm there are approximately 20 frequency transitions in each channel within a 5-sec stimulation period. Deutsch and Gregory (1978) have confirmed that no cerebral dominance effect is obtained when the octave stimuli are not part of long repetitive sequences of tones. The number of transitions in the Efron and Yund (1976) paradigm is much smaller. In view of the data presented by Halperin et al. (1973), the strong cerebral dominance effect achieved with the Deutsch octave stimuli may be due to the high number of transitions. If this is true, then the cerebral dominance effect evident in the octave illusion may depend more strongly on time variables than on pitch variables.

Rhythm Processing

Relatively little is known about the way in which dichotic time patterns are processed. Milner (1962) found a significant increase in error scores after right temporal lobectomy on the time test of the Seashore Measures of Musical Talents, which require the subject to judge the relative duration of two consecutive tones. Although this is not a dichotic listening task, Milner's data suggest that the right temporal lobe is involved in duration discrimination.

Spellacy (1970) reported that there was no preference for either ear when subjects responded to dichotic Morse code patterns. However, in his experiment the two channels were presented at different frequencies (1000 Hz and 1500 Hz) to maximize discrimination. It is possible that by introducing frequency differences between the ears, Spellacy's data reflect the complex interaction between time and frequency suggested by Gordon (1970). In other words, an opposing cerebral dominance effect for each of these variables might cancel the effects of both.

Robinson and Solomon (1974) presented right-handed subjects with dichotic rhythm patterns that consisted of from four to seven, short or long, sine wave pulses. The dichotic patterns were presented to subjects at matched frequencies, with simultaneous onset and offset of the two patterns. The data supported a hypothesis that rhythm is processed by the left hemisphere in most right-handed subjects. A similar hypothesis was satisfactorily defended by Gordon (1978), using rhythmic elements in dichotic melodies. This is interesting in view of the earlier model for dichotic listening proposed by Sparks et al. (1970). In this model competition between signals received by both ears occurs exclusively in the left hemisphere of left-dominant subjects.

An attempt to interpret the results of the rhythm experiments meets with a number of difficulties. One such problem concerns the population tested. Subjects used were right-handed or presumably left dominant for speech. There is no information that suggests how the left-handed population would respond. Throughout the behavioral literature, the idea is implicit that right-dominant subjects can be considered a reverse case of left-dominant

subjects. However, there is no concrete reason to assume that the right-dominant person processes information as does the left-dominant person, except with the directionality reversed. To the contrary, Corballis and Beale (1976) suggested that left-handedness, except in cases of pathology, is due to the cancellation of asymmetry, not its reversal. This implies that those persons normally described as right dominant should more accurately be classified as bidominant. If this is true, then information processing and related performance of the left-handed population cannot be inferred from experiments that use only right-handed subjects. One method of determining the exact contribution of asymmetrical function to the performance of the rhythm tasks might be to compare the responses of strongly dominant (right-handed) subjects to those of bidominant (left-handed) subjects. This has not been done.

In addition, many of the cerebral dominance effect experiments infer cerebral lateralization from responses that require the accurate lateralization of auditory stimuli. This is a questionable practice. The cues that are normally used to determine laterality in audition are phase and intensity differences between the stimuli to the two ears. In most experiments, intensities of the stimuli to the two ears have been matched. However, phase has not been controlled in any rhythm experiment that has been reported in the literature; in fact, it is impossible to match phase when using stimuli of different frequencies or verbal stimuli.

To summarize, there is no way to determine which aspects or amounts of reported laterality or favoritism are due to differences in the acoustical attributes of the stimuli arriving at the two ears and to determine how much laterality is due to differential cerebral processing of those stimuli. It would be easy to suggest that the right-ear preference of left-dominant subjects for rhythm may be due to some sort of bias toward the phase information presented to that ear. This would not undermine the validity of the results of experiments in the current literature; it does, however, illustrate the type of problem that is faced when one attempts to interpret these results.

Continued studies of rhythm may increase our understanding of many complex cognitive functions. The elements that constitute rhythm, such as sequence, order, and relative duration, are probably not limited to the auditory modality. Colavita (1977) supported a hypothesis that the insular temporal cortex in cats (which is equivalent to Wernicke's area in the dominant human hemisphere) is a polysensory association area with special importance for the perception of temporal patterns. Although he found no evidence of asymmetrical hemispheric function in the animals, he concluded that temporal pattern discrimination appears to represent the basis or antecedents of the higher order perceptual abilities in humans that are not specific to any single modality.

Neisser (1967; see summary, pp. 279–305) anticipated this hypothesis. He suggested that rhythm provides a framework to which verbal information can be attached and that such a framework serves both to integrate incoming information and to expedite recall. Investigation of such a hypothesis has proved difficult because of the logistical problems of measuring rhythm components in verbal information. In view of the more recent research, it may be that rhythm structure provides a framework not only for verbal material, as Neisser proposed, but also for the synthesis and analysis of all incoming perceptual information.

Conclusions

1. Pitch perception is probably a direct result of the frequency properties of the stimuli. Therefore, pitch sensation alone does not require differential cerebral processing. Only when some type of novel or complex time structure is generated in the stimulus presentation are the responses of subjects influenced by handedness or cerebral dominance.

2. The cerebral dominance effect evident in the results of the few existing rhythm experiments closely parallels that found in verbal experiments. Since adequate stimulus control is easier to achieve when using tonal stimuli than when using verbal stimuli, results of further rhythm experiments may add substantially to our knowledge of language. In

addition, recent reports suggest that rhythm structure may also provide a framework for the synthesis and analysis of all incoming perceptual information. If this should prove true, it is difficult to assess the effects it might have on our knowledge of human information processing.

3. A review of this area serves best to emphasize what we do not know. There is currently a trend in the literature toward separate investigation of pitch and rhythm variables. This is probably an excellent approach, although it may be difficult or impossible to separate them operationally. However, it should be possible to define the processing mechanisms peculiar to each variable through careful, controlled experimentation. Only when this elementary information is obtained can we hope to make any definitive statement about how they interact in complex stimuli such as speech signals. Renewed effort should produce sizeable increases in our knowledge in the near future.

References

Bryden, M. P. An evaluation of some models of laterality effects in dichotic listening. *Acta oto-Laryngologica*, 1967, *63*, 595–604.

Bryden, M. P. Binaural competition and division of attention as determinants of the laterality effects in dichotic listening. *Canadian Journal of Psychology*, 1969, *23*, 101–113.

Carpenter, M. B. *Human neuroanatomy*. Baltimore, Md.: Williams & Wilkins, 1976.

Christensen, I. P., & Gregory, A. H. Further study of an auditory illusion. *Nature*, 1977, *268*, 630.

Colavita, F. B. Theoretical review: Temporal pattern discrimination in the cat. *Physiology and Behavior*, 1977, *18*, 513–521.

Corballis, M. C., & Beale, I. L. *The psychology of left and right*. New York: Wiley, 1976.

Curry, F. K. W. A comparison of left-handed and right-handed subjects on verbal and nonverbal dichotic listening tasks. *Cortex*, 1967, *3*, 343–352.

Deutsch, D. An auditory illusion. *Nature*, 1974, *251*, 307–309.

Deutsch, D., & Gregory, A. H. Deutsch's octave illusion. *Nature*, 1978, *274*, 721.

Efron, R., Dennis, M., & Yund, E. W. The perception of dichotic chords by hemispherectomized subjects. *Brain and Language*, 1977, *4*, 537–549.

Efron, R., & Yund, E. W. Ear dominance and intensity independence in the perception of dichotic chords. *Journal of the Acoustical Society of America*, 1976, *59*, 889–898.

Geschwind, N. Language and the brain. *Scientific American*, 1972, *226*(4), 76–83.

Geschwind, N., & Levitsky, W. Human brain: Left-right asymmetries in temporal speech regions. *Science*, 1968, *161*, 186–187.

Goodglass, H., & Quadfasel, F. A. Language laterality in left-handed aphasics. *Brain*, 1954, *77*, 521–548.

Gordon, H. W. Hemispheric asymmetries in the perception of musical chords. *Cortex*, 1970, *6*, 387–398.

Gordon, H. W. Left hemisphere dominance for rhythmic elements in dichotically presented melodies. *Cortex*, 1978, *14*, 58–70.

Halperin, Y., Nachson, I., & Carmon, A. Shift of ear superiority in dichotic listening to temporally patterned nonverbal stimuli. *Journal of the Acoustical Society of America*, 1973, *53*, 46–49.

Jung, R. Summary of the conference. In V. B. Mountcastle (Ed.), *Interhemispheric relations and cerebral dominance*. Baltimore, Md.: Johns Hopkins University Press, 1962.

Kimura, D. Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology*, 1961, *15*, 166–171.

Kimura, D. Left-right differences in the perception of melodies. *Quarterly Journal of Experimental Psychology*, 1964, *16*, 355–358.

Kimura, D. Functional asymmetry of the brain in dichotic listening. *Cortex*, 1967, *3*, 163–178.

Milner, B. Laterality effects in audition. In V. B. Mountcastle (Ed.), *Interhemispheric relations and cerebral dominance*. Baltimore, Md.: Johns Hopkins University Press, 1962.

Milner, B. Taylor, L., & Sperry, R. W. Lateralized suppression of dichotically presented digits after commissural section in man. *Science*, 1968, *161*, 185–186.

Neisser, U. *Cognitive psychology*. New York: Appleton-Century-Crofts, 1967.

Nottebohm, F. Ontogeny of bird song. *Science*, 1970, *167*, 950–956.

Robinson, G. M., & Solomon, D. J. Rhythm is processed by the speech hemisphere. *Journal of Experimental Psychology*, 1974, *102*, 508–511.

Rossi, G. F., & Rosadini, G. Experimental analysis of cerebral dominance in man. In C. H. Millikan & F. L. Darley (Eds.), *Brain mechanisms underlying speech and language*. New York: Grune & Stratton, 1967.

Sparks, R., & Geschwind, N. Dichotic listening in man after section of neocortical commissures. *Cortex*, 1968, *4*, 3–16.

Sparks, R., Goodglass, H., & Nickel, B. Ipsilateral versus contralateral extinction in dichotic listening resulting from hemisphere lesions. *Cortex*, 1970, *6*, 249–260.

Spellacy, F. Lateral preferences in the identification of patterned stimuli. *Journal of the Acoustical Society of America*, 1970, *47*, 574–578.

Stevens, K. N., & House, A. S. Speech perception. In J. V. Tobias (Ed.), *Foundations of modern auditory theory* (Vol. 2). London: Academic Press, 1972.

Studdert-Kennedy, M., & Shankweiler, D. Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America*, 1970, *48*, 579–593.

Von Bonin, G. Anatomical asymmetries of the cerebral hemispheres. In V. B. Mountcastle (Ed.), *Interhemispheric relations and cerebral dominance*. Baltimore, Md.: Johns Hopkins University Press, 1962.

Yeni-Komshian, G., & Benson, D. A. Anatomical study of cerebral asymmetry in the temporal lobe of humans, chimpanzees, and rhesus monkeys. *Science*, 1976, *192*, 387–389.

Received July 31, 1978 ■

U.S. POSTAL SERVICE

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION

(Required by 39 U.S.C. 3685)

1. TITLE OF PUBLICATION: PSYCHOLOGICAL BULLETIN

A. PUBLICATION NO.: 9 3 1 1 4 0

2. DATE OF FILING: 9/30/79

3. FREQUENCY OF ISSUE: Bimonthly

A. NO. OF ISSUES PUBLISHED ANNUALLY: 6

B. ANNUAL SUBSCRIPTION PRICE: $15 member, $40 nonmember

4. LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, City, County, State and ZIP Code) (Not printers)*: 1400 N. Uhle Street, Arlington, VA 22201

5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*: 1200 17th Street N.W., Washington, D.C. 20036

6. NAMES AND COMPLETE ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER *(Name and Address)*: American Psychological Association, 1200 17th St., N.W., Wash., D.C. 20036

EDITOR *(Name and Address)*: R.J. Herrnstein, Harvard University, Cambridge, MA 02138

MANAGING EDITOR *(Name and Address)*: Anita DeVivo, 1400 N. Uhle St., Arlington, VA 22201

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)*

NAME	ADDRESS
APA	1200 17th St. N.W., Wash. D.C. 20036

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

NAME	ADDRESS
None	

9. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES *(Section 132.122, PSM)*
The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes *(Check one)*

XX HAVE NOT CHANGED DURING PRECEDING 12 MONTHS ☐ HAVE CHANGED DURING PRECEDING 12 MONTHS *(If changed, publisher must submit explanation of change with this statement.)*

10. EXTENT AND NATURE OF CIRCULATION	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NO. COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
	11,258	11,209
A. TOTAL NO. COPIES PRINTED *(Net Press Run)*	- 0 -	- 0 -
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES	9,385	9,721
2. MAIL SUBSCRIPTIONS	9,385	9,721
C. TOTAL PAID CIRCULATION *(Sum of 10B1 and 10B2)*	173	186
D. FREE DISTRIBUTION BY MAIL, CARRIER OR OTHER MEANS SAMPLES, COMPLIMENTARY, AND OTHER FREE COPIES	9,558	9,907
E. TOTAL DISTRIBUTION *(Sum of C and D)*	1,700	1,302
F. COPIES NOT DISTRIBUTED 1. OFFICE USE, LEFT OVER, UNACCOUNTED, SPOILED AFTER PRINTING	- 0 -	- 0 -
2. RETURNS FROM NEWS AGENTS	11,258	11,209
G. TOTAL *(Sum of E, F1 and 2—should equal net press run shown in A)*		

11. I certify that the statements made by me above are correct and complete. ▶ SIGNATURE AND TITLE OF EDITOR, PUBLISHER, BUSINESS MANAGER, OR OWNER

12. FOR COMPLETION BY PUBLISHERS MAILING AT THE REGULAR RATES *(Section 132.121, Postal Service Manual)*

39 U. S. C. 3626 provides in pertinent part: "No person who would have been entitled to mail matter under former section 4359 of this title shall mail such matter at the rates provided under this subsection unless he files annually with the Postal Service a written request for permission to mail matter at such rates."

In accordance with the provisions of this statute, I hereby request permission to mail the publication named in Item 1 at the phased postage rates presently authorized by 39 U. S. C. 3626.

SIGNATURE AND TITLE OF EDITOR, PUBLISHER, BUSINESS MANAGER, OR OWNER

(See instructions on reverse)

PS Form 3526 *(Page 1)* Mar. 1977

Psychological Bulletin
1979, Vol. 86, No. 6, 1350–1352

Practical Considerations in Choosing a MANOVA Test Statistic: A Rejoinder to Stevens

Chester L. Olson
Camrose Lutheran College, Alberta, Canada

The recommendation of the Pillai-Bartlett trace V for general use as the test statistic in multivariate analysis of variance (MANOVA) is reexamined and reaffirmed in light of criticisms by Stevens. Empirical data are presented to show that the V test's substantial robustness advantage does not require extreme violations of assumptions, and examples are cited to show the occurrence in practice of the type of heteroscedasticity in which the V test particularly excels and the occurrence of the pattern of population mean differences for which the V test is more powerful than its rivals.

In an earlier article (Olson, 1976), I presented evidence from power and robustness studies that the Pillai-Bartlett trace statistic V should be favored for general use as the test statistic in multivariate analysis of variance (MANOVA) and that Wilks's likelihood ratio W and the Hotelling-Lawley trace T could be considered equivalent to V in very large samples. Now Stevens (1979) has argued (a) that except in certain extreme cases, T and W are nearly as robust as V, (b) that such exceptions occur very infrequently in real data, and (c) that the V test should be used when there are population mean differences in several canonical dimensions of the multivariate space, but any of the V, W, or T statistics may be used when the population differences are concentrated in one dimension. The following discussion of these arguments serves to clarify some practical considerations in choosing a MANOVA test statistic and to underscore my earlier recommendation.

First, concerning the first argument: Is it only in extreme cases that V's robustness advantage is large enough to make much difference in practice? Stevens (1979, Table 1) shows empirical Type I error rates for 45 examples of heteroscedasticity from Olson (1973). When the contamination that produced the excessive variance in the one anomalous group was concentrated $[\mathbf{D} = \mathbf{C}(d)]$ in a single canonical variate, error rates for T, W, and V remained about equally close to the nominal level, although V was most often closest. When the anomalous group came from a population more variable in all dimensions $(\mathbf{D} = d\mathbf{I})$, the V test was almost always disturbed less than its rivals. Exploring the notion that V generally performed no more than 1.5 and 2.5 percentage points better than W and T, respectively, Stevens found that "eight of the nine cases in Table 1 in which the differences in error rates are larger. . . correspond to very large subgroup variance differences on all variables (36I)" (p. 355). There are two major reasons for this finding.

1. Stevens omitted some of Olson's (1973) examples of heteroscedasticity of the nonconcentrated type $(\mathbf{D} = d\mathbf{I})$. Table 1 of the present article gives Type I error rates for W and V in all the omitted examples for $d = 4$, 9, and 36. V bettered W by Stevens's criterion of at least 1.5 percentage points in 21 of the 26 cases, and of the 5 remaining cases, 3 would not be expected to show a difference because the sample sizes were very large relative to the number of variables (Olson, 1976, p. 583). Note that V surpassed Stevens's criterion even in cases in which the subgroup variance differences were not extreme ($d = 4$ and 9).

2. Olson (1973) examined far more cases with $d = 36$ than with any other value of d.

Requests for reprints should be sent to Chester L. Olson, Department of Psychology, Camrose Lutheran College, Camrose, Alberta, Canada T4V 2R3.

Table 1
Empirical Type I Error Rates for W *and* V *at the Nominal .05 Level When Variances in One Group Were* d *Times the Variances in the Other Groups*

			$d = 4$			$d = 9$			$d = 36$		
No. of variables	No. of groups	Group size	W	V	Difference[a]	W	V	Difference[a]	W	V	Difference[a]
2	6	50							163	162	001[b]
2	10	5							235	189	046[c]
2	10	10							208	194	014
3	3	5	080	067	013	134	093	041[c]	245	163	082[c]
3	3	50							098	094	004[b]
3	6	5	088	070	018[c]	162	100	062[c]	289	162	127[c]
3	6	10	092	077	015[c]	146	124	022[c]	224	186	038[c]
3	6	50							173	164	009[b]
3	10	5							307	167	140[c]
3	10	10							251	202	049[c]
6	3	5							489	092	397[c]
6	6	5							532	047	485[c]
6	6	10							361	177	184[c]
6	10	5							517	053	464[c]
6	10	10							356	171	185[c]
10	3	5							451	183	268[c]
10	6	5							707	073	634[c]
10	6	10							580	098	482[c]
10	10	5							732	051	681[c]
10	10	10							529	108	421[c]

Note. Data from Olson (1973). The decimal point preceding each digit triplet has been omitted; for example, 080 denotes .080.
[a] The standard error of such differences is approximately .007.
[b] W and V were almost equivalent because of the very large sample size.
[c] Differences of .015 or more.

The reader can verify from the data in Table 1 (and in Stevens's Table 1) that as d increased from 1 (at which point the actual error rate equals the nominal level) to 4 to 9 to 36, the actual error rate increased regularly, with a tendency to increase proportionately more in the range from 1 to 9 than in the range from 9 to 36. Thus one can get a fairly good idea of error rates and differences between error rates at intermediate values of d by interpolation between $d = 1$ and $d = 36$. Any reasonable interpolation in Table 1 shows that d does not have to be nearly as extreme as 36 for V to be markedly superior to W for small to moderate sample sizes.

Let us turn now to the second argument: Is it true that the d**I** pattern of heteroscedasticity, in which one population is more variable than the others in all dimensions, occurs very infrequently in practice? To support an affirmative answer, Stevens summarized real data from nine investigations. (Actually, one of the nine is not quite real: Smith, Gnanadesikan, and Hughes, 1962, arbitrarily divided their data into the four groups "just for illustrative purposes" [p. 28].) Of course, with real (sample) data, one has no way of being sure what the underlying populations looked like, but even in Stevens's intended counterexamples, there is evidence of the type of heteroscedasticity in question: Wright's (1975) first group was more variable on all nine subtests than the other three groups, which had approximately equal variances.[1] This example suggests the d**I** pattern of heteroscedasticity with $d \approx 2.2$. In Meichenbaum's (1975) data from the three subscales of the consequences test, variances were largest for the self-instructional group but unequal for the other two

[1] Note that the variance on the spelling subtest for girls in a traditional school was erroneously reported in Stevens's article as 2.25 rather than 1.82.

groups, as noted by Stevens. Data not reported by Stevens from the same experiment revealed that the self-instructional group was more variable on all three subscales of the unusual uses test than the other two groups, which were about equally variable, suggesting non-concentrated heteroscedasticity with $d \approx 1.9$. The values of d are small, but these examples illustrate that the $d\mathbf{I}$ pattern of heteroscedasticity is entirely realistic.

Finally, concerning the third argument: What recommendation can be made to guide the MANOVA user? Stevens correctly noted that in terms of power, the preferred test depends on whether the noncentrality, or inequality of population means, is (a) heavily concentrated in one canonical variate, in which case the T and W tests are more powerful than V, especially with more than five or six variables, or (b) more diffuse, appearing in several dimensions, in which case the V test is more powerful even if the noncentrality is not distributed equally across the several dimensions (Olson, 1973; Schatzoff, 1966). Nor can one refute Stevens's inference that concentrated noncentrality is probably common in behavioral research. However, there is also evidence that moderately diffuse noncentrality is by no means uncommon in practice (e.g., Bock, 1975, p. 408; Jones, 1966, pp. 258–265; Overall & Klett, 1972, pp. 285–292), and since noncentrality is a property of the population, not of the sample, the experimenter has no way of knowing for certain which form of noncentrality, if any, exists in a particular population of interest. If significance levels are to mean anything, it is certainly not appropriate to choose one's test statistic on the basis of the sample outcome that is to be tested! Thus the experimenter, faced with real data of unknown noncentrality and trying to follow Stevens's recommendation to use V for diffuse noncentrality and any of the V, W, or T statistics for concentrated noncentrality, must always choose V, a choice compatible with my original recommendation (Olson, 1976, p. 585).

In summary, the V test is sometimes more powerful than W or T and sometimes less powerful, but it is consistently more robust, sometimes by a substantial margin, and all of these situations are realistic possibilities in practice. It must be noted that the conclusion to be drawn from these facts depends on one's relative distaste for Type I and Type II errors, and my own preference is to ensure that the Type I error rate remains close to the nominal level. Therefore, the V test is recommended for routine use.

References

Bock, R. D. *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill, 1975.

Jones, L. V. Analysis of variance in its multivariate developments. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.

Meichenbaum, D. Enhancing creativity by modifying what subjects say to themselves. *American Educational Research Journal*, 1975, *12*, 129–145.

Olson, C. L. *A Monte Carlo investigation of the robustness of multivariate analysis of variance*. Unpublished doctoral dissertation, University of Toronto, Canada, 1973.

Olson, C. L. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 1976, *83*, 579–586.

Overall, J. E., & Klett, C. J. *Applied multivariate analysis*. New York: McGraw-Hill, 1972.

Schatzoff, M. Sensitivity comparisons among tests of the general linear hypothesis. *Journal of the American Statistical Association*, 1966, *61*, 415–435.

Smith, H., Gnanadesikan, R., & Hughes, J. B. Multivariate analysis of variance (MANOVA). *Biometrics*, 1962, *18*, 22–41.

Stevens, J. P. Comment on Olson: Choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 1979, *86*, 355–360.

Wright, R. J. The affective and cognitive consequences of an open education elementary school. *American Educational Research Journal*, 1975, *12*, 449–468.

Received January 15, 1979 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1353–1373

Index of Literature Reviews and Summaries in the *Psychological Bulletin*, 1967–1978

Ludy T. Benjamin, Jr., and Leigh S. Shaffer
Nebraska Wesleyan University

This index updates a previous compilation by Thomas Andrews and Frances Kerr, which was published in 1967 in the *Psychological Bulletin* (*68*, 178–212), covering the years 1940–1966. The *Psychological Bulletin* has continued to be the major source of literature reviews and summaries on topics of interest to psychologists, and the present index was designed to facilitate access to that material. Like the earlier index, the current one includes all articles that were wholly or largely reviews of a specialized area of literature. Published comments on those review articles have also been included. Those articles that were judged not to qualify as literature reviews were excluded from this index. The author index is completely cross-referenced. Articles are designated by author, volume number, and page number(s). In the subject index, the number of items in the reference section of each article is also provided. The letter *P* following a subject listing indicates that the particular subject is only part of the review article. In preparing the subject index the articles were cross-referenced on a large number of headings in an attempt to make topical retrieval by the researcher as fruitful as possible.

Key to Volume Numbers by Year

67–1967	*74*–1970	*81*–1974
68–1967	*75*–1971	*82*–1975
69–1968	*76*–1971	*83*–1976
70–1968	*77*–1972	*84*–1977
71–1969	*78*–1972	*85*–1978
72–1969	*79*–1973	
73–1970	*80*–1973	

Author Index

Requests for reprints should be sent to Ludy T. Benjamin, Jr., who is now at the Educational Affairs Office, American Psychological Association, 1200 17th Street, N.W., Washington, D.C. 20036.

Subject Index

Received June 16, 1978 ■

Psychological Bulletin
1979, Vol. 86, No. 6, 1374–1384

Index of Reviews and Notes on Statistical Methods and Research Design in the *Psychological Bulletin*, 1967–1978

Leigh S. Shaffer and Ludy T. Benjamin, Jr.
Nebraska Wesleyan University

This index updates the previous index of reviews and notes on statistical methods and research design compiled in 1967 by Thomas Andrews and published by the *Psychological Bulletin* (*68*, 213–220). This index complements the index of literature reviews and summaries by including useful treatments of statistical procedures and research design considerations that did not qualify as literature reviews. Organization of this index is identical to that of the index of literature reviews and summaries and articles on topics of statistics and research design that are included in the larger index are also included in this index.

Author Index

Requests for reprints should be sent to Leigh S. Shaffer, Department of Psychology, Nebraska Wesleyan University, Lincoln, Nebraska 68504.

Subject Index

Received June 16, 1978 ■

Psychological Bulletin

R. J. Herrnstein, Editor
Harvard University

Gene V Glass, Associate Editor
University of Colorado

VOLUME 86 1979

Published Bimonthly by the American Psychological Association, Inc.
1200 Seventeenth Street, N.W., Washington, D.C. 20036

Editorial Consultants

Journal Staff

Anita DeVivo
Executive Editor

Ann I. Mahoney
Manager, Journal Production

Barbara R. Richman
Production Supervisor

Michal M. Keeley
Production Editor

Robert J. Hayward
Advertising Representative

Juanita Brodie
Subscription Manager

Author Index to Volume 86

Key to Pagination

Issue No.	*Month*	*Pages*	*Issue No.*	*Month*	*Pages*
1	January	1–215	4	July	643–888
2	March	217–428	5	September	889–1168
3	May	429–641	6	November	1169–1384

ARTICLES

OTHER

CONTENTS (continued)

Miller Appointed Editor, 1981–1986

The Publications and Communications Board of the American Psychological Association announces the appointment of George A. Miller as editor of *Psychological Bulletin* for the years 1981–1986. As of January 1, 1980, manuscripts should be directed to the Editor-elect:

George A. Miller
Department of Psychology
Princeton University
Princeton, New Jersey 08540

A.M

29/3/80